CHEMICAL
RESEARCH,
DEVELOPMENT &
ENGINEERING
CENTER

CRDEC-CR-88006

THE INVERSE PROBLEM AND THE PSEUDO-EMPIRICAL ORTHOGONAL FUNCTION METHOD OF SOLUTION - PART 1: THEORY; PART 2: APPLICATION

by Avishai Ben-David
Benjamin M. Herman
John A. Reagan

University of Arizona
Tucson, AZ 85721

November 1987

U.S. ARMY
ARMAMENT
MUNITIONS
CHEMICAL COMMAND

Aberdeen Proving Ground, Maryland 21010-5423

UNCLASSIFIED
RITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b RESTRICTIVE MARKINGS
SECURITY CLASSIFICATION AUTHORITY	3 DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
DECLASSIFICATION / DOWNGRADING SCHEDULE	
PERFORMING ORGANIZATION REPORT NUMBER(S) CRDEC-CR-88006	5 MONITORING ORGANIZATION REPORT NUMBER(S)
NAME OF PERFORMING ORGANIZATION University of Arizona	6b OFFICE SYMBOL (If applicable)
	7a NAME OF MONITORING ORGANIZATION
ADDRESS (City, State, and ZIP Code) Tucson, AZ 85721	7b ADDRESS (City, State, and ZIP Code)
NAME OF FUNDING / SPONSORING ORGANIZATION CRDEC	8b OFFICE SYMBOL (If applicable) SMCCR-RSP-B
	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER DAAK11-84-K0014
ADDRESS (City, State, and ZIP Code)	10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO	TASK NO.	WORK UNIT ACCESSION NO

TITLE (Include Security Classification)
The Inverse Problem and the Pseudo-Empirical Orthogonal Function Method of Solution - Part 1: Theory; Part 2: Application

PERSONAL AUTHOR(S)
Ben-David, Avishai; Herman, Benjamin M.; and Reagan, John A.

TYPE OF REPORT	13b TIME COVERED	14 DATE OF REPORT (Year, Month, Day)	15 PAGE COUNT
Contractor	FROM 84 Aug TO 86 Dec	1987 November	78

SUPPLEMENTARY NOTATION
COR: Dr. Jerald Bottiger, SMCCR-RSP-B, (301) 671-2395

COSATI CODES		
FIELD	GROUP	SUB-GROUP
20	06	

18 SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Inverse Problem Orthogonal Function Method
Light Scattering

ABSTRACT (Continue on reverse if necessary and identify by block number)

—A library of mathematical functions or set of observations is used to form a set of orthonormal basis functions. It is assumed that any unknown solution can be constructed from a linear sum of these basis functions. A solution with a smoothing constraint and/or positivity constraints is developed. An analysis of the information contained in the measurements about the unknown solution is given.

The amount of information (types of solutions, accuracy, and moments of the solution) about tropospheric rural aerosol size distribution that can in theory be obtained from backscattered measurements, without using any additional information about the anticipated assumptions (constraints) must be used to solve for aerosol size distribution. The inferred solution reflects assumptions and is therefore not objective. The quality of the solution depends on the applicability of the constraints to the given problem. Solutions for the inverse problem using- (continued on reverse)

DISTRIBUTION / AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED

NAME OF RESPONSIBLE INDIVIDUAL	22b TELEPHONE (Include Area Code)	22c OFFICE SYMBOL
SANDRA J. JOHNSON	(301) 671-2914	SMCCR-SPS-T

FORM 1473, 84 MAR 83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. ABSTRACT (continued)

the pseudo-empirical orthogonal method of solution were obtained using two types of constraints: positivity and positivity combined with smoothing. Results of the method, when backscattered radiation is used as measurements, are presented. Discussion on the limitations of the method and the effects upon the solution of the different assumptions that are used is given. Some possible uses of the solution are considered.

PREFACE

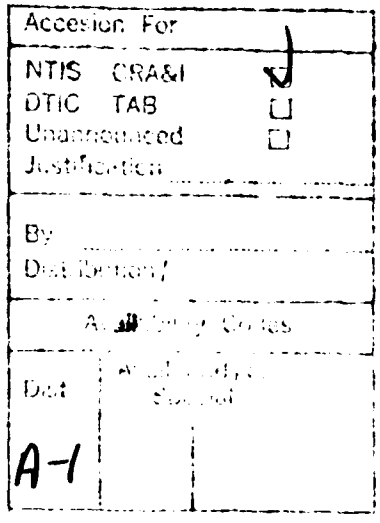
The work described in this report was authorized under Contract No. DAAK11-84-K-0014. This work was started in August 1984 and completed in December 1986.

The use of trade names or manufacturers' names in this report does not constitute an official endorsement of any commercial products. This report may not be cited for purposes of advertisement.

This report has been approved for release to the public.

Acknowledgment

Editing of the manuscript by Margaret Sanderson Rae is gratefully acknowledged.



Blank

CONTENTS

LIST OF FIGURES

THE INVERSE PROBLEM AND THE PSEUDO-EMPIRICAL ORTHOGONAL FUNCTION METHOD OF SOLUTION PART 1: THEORY; PART 2: APPLICATION

1. THEORY

1.1 Introduction

Many schemes have been developed to solve the so-called "inverse problem." However, it remains a fact, regardless of the scheme, that normally the information content in a given set of measurements is severely limited. Therefore, our recoverable knowledge of the unknown, if deduced solely from the measurements, is also going to be severely limited. The difference between the various inversion schemes is primarily due to the additional information that the set of equations is given. This additional information is normally in the form of "physically plausible" constraints.

The list of methods is too long to repeat here [the interested reader can be referred to Twomey,[1] Deirmendjian,[2] and Bottiger[3]], but it follows that, in any method of solution employing one or more constraints, the final solution will depend to some degree on the validity of the constraint for the particular problem and, therefore, is not completely objective.

Given a set of measurements, g_i, $i = 1, 2, \cdots m$, the governing equation for the inverse problem can normally be written as a Fredholm integral of the first kind:

$$g_i = \int_a^b k_i(x)\, f(x)\, dx \qquad (1)$$

where $k_i(x)$ is the i^{th} kernel of the problem, and $f(x)$ is the unknown. The measurement g_i is, therefore, a dot product in the function space between the i^{th} kernel and $f(x)$. Therefore, the part of the solution, $f(x)$, that can be recovered from the measurements must lie within the function space spanned by the kernels. Any solution or component of

a solution which is outside this function space requires additional information (constraints) and/or assumptions, in order to be recovered. The presence of measurement errors and uncertainties in the mathematical model and the physical processes represented by the kernels act to further increase the unrecoverable part of the solution.

Equation (1) can be approximated in some fashion by a discrete sum, such that f(x) is calculated at some set of x's, and therefore may be rewritten in a matrix form:

$$\vec{g} = \tilde{A}\,\vec{f} \tag{2}$$

where $\vec{g}$ is an m×1 vector made up of the set of measurements; $\tilde{A}$ is an m×n matrix representing the kernel and may contain weighting factors which depend on the quadrature formula used for converting from an integral to a finite sum; and $\vec{f}$ is the unknown n×1 column vector whose elements are f_i. The direct solution of the basic Eq. (2) is

$$\vec{f} = \tilde{A}^{-1}\vec{g}$$

or

$$\vec{f} = (\tilde{A}^T\tilde{A})^{-1}\,(\tilde{A}^T\vec{g}), \tag{3}$$

when the $\tilde{A}$ matrix is not square where T denotes the transposed matrix. All methods of solution for Eq. (2) require the computation (directly or indirectly) of the inverse of the matrix $\tilde{A}^T\tilde{A}$.[4] The instability as a result of the inverse operation and its relation to the eigenvalues and eigenvectors of the matrix $\tilde{A}^T\tilde{A}$ will be demonstrated as follows.[5,6]

Let the exact equation to be solved be written as:

$$\vec{b} = \tilde{G}\vec{f} \tag{4}$$

where $\vec{b}$ is the transposed measurement vector $\tilde{A}^T\vec{g}$ and $\tilde{G}$ is the $\tilde{A}^T\tilde{A}$

matrix obtained from Eq. (2) after premultiplication by $\tilde{A}^T$. If a small perturbation $\delta\vec{b}$ is assumed in the measurements and the resulting perturbation $\delta\vec{f}$ is estimated such that

$$\vec{b} + \delta\vec{b} = \tilde{G}(\vec{f} + \delta\vec{f}), \tag{5}$$

then subtracting Eq. (4) from Eq. (5) yields

$$\begin{aligned} \delta\vec{b} &= \tilde{G}\delta\vec{f} \\ \delta\vec{f} &= \tilde{G}^{-1}\delta\vec{b}. \end{aligned} \tag{6}$$

The matrix $\tilde{G}$ is a symmetric and real matrix, so that its eigenvalues λ_i are real and positive and $0 < \lambda_1 < \lambda_2 \cdots < \lambda_n$. The eigenvalues of $\tilde{G}^{-1}$ are $0 < 1/\lambda_n < 1/\lambda_{n-1} \cdots 1/\lambda_1$, and the eigenvectors of $\tilde{G}$ and $\tilde{G}^{-1}$ are $\vec{u}_1, \vec{u}_2 \cdots \vec{u}_n$. Assume now that any perturbation vector $\delta\vec{b}$ can be written as a linear combination of the eigenvectors $\vec{u}_i$, $i = 1,2 \cdots n$. For the specific case for which $\delta\vec{b} = \varepsilon\vec{u}_i$, where ε is the magnitude of $\delta\vec{b}$, Eq. (6) may be written as:

$$\delta\vec{f} = \tilde{G}^{-1}\varepsilon\vec{u}_i. \tag{7}$$

By using the eigenvalue equation for $\tilde{G}^{-1}$, $\tilde{G}^{-1}\vec{u}_i = (1/\lambda_i)\vec{u}_i$ yields:

$$\delta\vec{f} = \frac{1}{\lambda_i}\,\delta\vec{b}. \tag{8}$$

The error of magnitude $|\delta\vec{b}|$ is amplified by a factor $1/\lambda_i$, which can be very large when λ_i is close to zero for a nearly singular matrix. Furthermore, from Eq. (4), it follows that $\vec{f} = \tilde{G}^{-1}\vec{b}$, and for $\vec{b} = \alpha\vec{u}_n$, where α is the magnitude of $\vec{b}$, the minimum magnitude for $\vec{f}$ will be

$$\vec{f} \geq \frac{1}{\lambda_n}\,\vec{b}, \tag{9}$$

where λ_n is the largest eigenvalue.

Combining Eqs. (8) and (9) yields an estimate for the relative error:

$$\left(\frac{|\delta\vec{f}|}{|\vec{f}|}\right) \leq \left(\frac{\lambda_n}{\lambda_1}\ \frac{|\delta\vec{b}|}{|\vec{b}|}\right), \tag{10}$$

which says that the relative error in $\vec{f}$ is less than or equal to the relative error in $\vec{b}$ times the ratio of the largest to the smallest eigenvalue of G.

By improving the quality of the measurements and the model, the magnitude of the error $|\delta\vec{b}| = \varepsilon$ can be controlled. However, it is important to note that the direction of $\delta\vec{b}$ is outside of control. The probability always exists that some of the error vector will be in the direction of $\vec{u}_1$, and a very large error in the solution $\vec{f}$ will result.

If it is desired to include in the error magnification analysis the combined effect of error in the measurements and error in the kernel function, Twomey[1] (pp. 207 - 210) can be followed to solve for the case in which $\vec{\varepsilon}$ is an error measurement vector that obeys normal additive statistics. $\varepsilon_i = N(0,e^2g_i^2)$; N represents normal statistics, and e is a fraction error in g_i. Letting $\eta_i(r)$ be the error in the kernel $k_i(r)$, which obeys normal additive statistics, $\eta_i(r) = N[0,P^2k_i^2(r)]$, where P is a fraction error in $k_i(r)$ and there is no correlation between $\vec{\varepsilon}$ and $\vec{\eta}$. The worst case of relative error magnification can be determined if the problem is normalized such that

$$\int_a^b f^2(x)\,dr = 1, \text{ and } \int_a^b k_i^2(x)\,dx = 1,\ i = 1,2 \cdots m \text{ to be}$$

$$\frac{|\delta(\vec{f})|^2/|\vec{f}|^2}{[|\vec{\varepsilon}|^2/|\vec{g}|^2]} = \frac{\lambda_n}{\lambda_1} + \frac{P^2\lambda_n}{e^2\lambda_1^2} + P^2\frac{\lambda_n}{\lambda_1}, \tag{11}$$

or the best case of relative error magnification is expressed by:

$$\frac{|\delta(\vec{f})|^2/|\vec{f}|^2}{[|\vec{\varepsilon}|^2/|\vec{g}|^2]} = \frac{\lambda_n}{\lambda_1} + \frac{P^2}{e^2\lambda_1} + P^2\frac{\lambda_1}{\lambda_n} . \qquad (12)$$

Both Eqs. (11) and (12) give Twomey's results when $P \to 0$ (no error in the kernels).

The problems of more unknowns than available independent equations and of inherent instability arise when a solution of the basic equation, Eq. (2), is attempted. It is important to note that, even when the number of measurements equals or exceeds the number of unknowns, due to errors, the resulting equations are not all independent. To supply more independent equations, relationships betweer the unknown solution points can be assumed or assumptions about other properties of the expected solution can be made.

If no constraints are available or if those available are insufficient, no solution with finite error bounds can be found.[7] The additional assumptions (constraints) can be mathematical (weighting functions) or physical (additional assumed physical characteristics). In some cases, the constraints can be hidden. As an example, Smith[8] required that the solution be a linear combination of the weighting functions, and Chanine[9,10] used linear interpolation between the unknown solution points (temperature at different elevations) as noted by Rodgers.[7] The constraints can be explicitly stated, as did Twomey,[11,12] Fleming and Wark,[13] Wark and Fleming,[14] Herman *et al.*,[15] and others.

The applicability of the constraint to a given problem is very important because the additional set of constraint equations serves

as virtual measurements. If the proposed constraints do not properly describe the physics of the problem, to use them would be tantamount to adding measurements with large errors. For example, if on a particular day the aerosol size distribution was a very erratic function and a smoothing constraint was applied, the result would be two sets of contradictory equations. The quality and type of the constraint applied are reflected in the solution and make it nonobjective.

Frequently the solution, f(x), is used to calculate different properties of f(x) [*i.e.*, moments of f(x)] or other derived properties as an input for other models. Therefore, it is important to know how sensitive the measurements are to these computed properties and the standard deviation of the solution. If the measurements are not sensitive to these properties, then these computed properties are not necessarily appropriate for use in modelling. Analysis of the kernels can give information about types of solutions and properties of the solutions to which the measurements are sensitive, as will be shown later.

For those inverse problems for which there exists a large body of observed solutions [distribution functions f(x)], such as temperature soundings, ozone vertical distributions, etc., one can make use of this large reservoir of observed "solutions" using the method of empirical orthogonal functions.[16-19] In this method, the observed information is put into the form of a matrix from which the eigenvectors are determined. From these eigenvectors, a set of orthonormal basis functions may be constructed, which are used as additional information to constrain the unknown solution to be composed of a linear combination of these observed distributions. It is inherently assumed in this method that the orthonormal basis functions form a complete set.

For those problems for which we do not have any reliable library of observed "solutions," *e.g.*, when f(x) is an aerosol size distribution, the basis functions can be constructed from a library of mathematical functions. Although these mathematical functions do not describe existing functions f(x), they can produce orthonormal basis functions from which many types of anticipated solutions can be constructed. These orthonormal basis functions will be called pseudo-empirical orthogonal functions, as their source is from a library of mathematical functions. Inherently assumed in the method is that the library of assumed functions can be used in linear combinations to yield any real function which describes the unknown f(x).

The approach of expanding an unknown function as a linear combination of orthonormal basis functions can be found in many mathematical reference books.[20,21] In this work, this approach will be applied for solving the inverse problem.

1.2 Method of Solution

It is assumed that the unknown solution function f(x) can be constructed from a linear sum of orthonormal basis functions $\phi(x)$ with coefficients a_i, i = 1,2 ••• p,

$$f(x) = \sum_{i=1}^{p} a_i \phi_i(x) \quad . \tag{13}$$

The validity of this assumption will be discussed further on. The coefficient set a_i can be represented in matrix notation by a column vector $\vec{a}(p\times1)$ whose elements are a_i. The basis functions $\phi_i(x)$, i = 1,2 ••• p, can be written in matrix notation as a column vector $\vec{\phi}(x)$, (p×1), where each element is $\phi_i(x)$ and x is a continuous variable. If

x is discretized (x_1 to x_n), the basis functions can be written as a matrix $\tilde{\phi}(p \times n)$, while $\vec{f}$ is, as before, a column vector $(n \times 1)$. By using the above matrix notation, the unknown solution can be written as:

$$\vec{f} = \tilde{\phi}^T \vec{a}. \tag{14}$$

Substitution of the expansion for the unknown f(x) [Eq. (13)] into the basic Eq. (1) yields:

$$g_i = \int_a^b k_i(x) \sum_{j=1}^{p} a_j \phi_j(x)\, dx = \sum_{j=1}^{p} a_j \int_a^b k_i(x) \phi_j(x)\, dx \tag{15}$$

or in matrix notation:

$$\vec{g} = \tilde{A}\vec{a}\ , \tag{16}$$

where now the $\tilde{A}$ matrix elements are composed of the various terms of the integral on the right-hand side of Eq. (15), which are inner products between the physical kernel of the problem $k_i(x)$ and the mathematical basis functions $\phi_j(x)$, which express assumed prior knowledge about the unknown function f(x) such that

$$A_{ij} = \int_a^b k_i(x) \phi_j(x)\, dx\ . \tag{17}$$

The condition number (the ratio between the largest and smallest eigenvalues of $\tilde{A}^T\tilde{A}$) is generally large and, as a result, the solution, $\vec{a} = (\tilde{A}^T\tilde{A})^{-1}\tilde{A}^T\vec{g}$ and $\vec{f} = \tilde{\phi}^T\vec{a}$, is unstable. This results from the fact that the number of unknown coefficients (a_i) which are needed for the approximation, Eqs. (13) and (15), is usually much larger than the limited number of independent measurements. As a result, it becomes

necessary to employ additional sets of equations in the form of constraints that express properties anticipated in the expected solution.

The first constraint employed here is a smoothing constraint, since f(x) is expected to be smooth for most low-resolution measurements of g_i. A smoothing constraint will also serve as a low-pass filter that will reject high-frequency oscillations. Mathematically, high-frequency components usually are nearly orthogonal to the kernels and are therefore undesirable because they do not contribute significantly to the measurements.

A second-order derivative of the solution with respect to x will be used as a measure of smoothness, and the constraint applied will be to minimize the sums of the squares of the second derivatives. In matrix notation, the smoothing operation can be written as:

$$\frac{\partial^2 f(x)}{\partial x^2} = \tilde{S}\,\tilde{\phi}^T\vec{a}$$

and

$$\left[\sum_i \frac{\partial^2 f(x_i)}{\partial x_i^2}\right]^2 = \vec{a}^T\tilde{\phi}\tilde{S}^T\tilde{S}\tilde{\phi}^T\vec{a} = q_2 \ , \qquad (18)$$

where $\tilde{S}$ is an operator (matrix), as derived by Twomey,[1] which, when applied to $\vec{f} = \tilde{\phi}^T\vec{a}$, yields the second derivatives at each point. Letting the matrix $\tilde{\phi}\tilde{S}^T\tilde{S}\tilde{\phi}^T = \tilde{H}_S$, the second-order smoothing matrix, Eq. (18) can be rewritten as:

$$q_2 = \vec{a}^T\tilde{H}_S\vec{a} \qquad (19)$$

where q_2 is a scalar parameter indicating the degree of smoothness of the solution, f(x). As the numerical value of q_2 becomes smaller, the solution becomes smoother.

The second constraint to be employed is the condition of non-negative solution points. Unlike the property of smoothness, which is present in most but not all cases, positivity is a property that must be observed at all times (*e.g.*, when f(x) is the aerosol size distribution: temperature profile, ozone amount, etc.).

The positivity property can be formulated if a function q_3 is created such that

$$q_3(x_j) = f^2(x_j) - y_j f(x_j) \qquad y_j > 0 \tag{20}$$

where y_j is the j^{th} positive component of a suitably chosen vector to be described later. Inspection of Eq. (20) and the graph in Fig. 1 shows that the function $q_3(x_j)$ results in a small magnitude for positive $f(x_j)$ and a large magnitude for negative $f(x_j)$. The slope for negative values of $f(x_j)$ is much steeper than is the slope for positive $f(x_j)$. The function $q_3(x_j)$ produces a minimum (a single minimum) for a positive value of $f(x_j) = y_j/2$. The constraint of minimizing the function $q_3(x_j)$ will push the solution toward $y_j/2$, a positive number since y_j was clearly positive. Equation (20) can be written in matrix form as:

$$q_3(x_j) = \vec{a}^T \tilde{\phi} \vec{E}_j \vec{E}_j^T \tilde{\phi}^T \vec{a} - y_j \vec{a}^T \tilde{\phi} \vec{E}_j \tag{21}$$

where $\vec{E}_j$ is a column vector (n×1) in which all the elements are zero, except for one in the j^{th} element. Equation (20) is true for any point x_j so that there are n equations represented by Eq. (20) for all points x_1 to x_n.

The Euler-Lagrange method will be employed to solve the problem where:

$$|\tilde{A}\vec{a} - \vec{g}|^2 \leq q_1 \ , \tag{22}$$

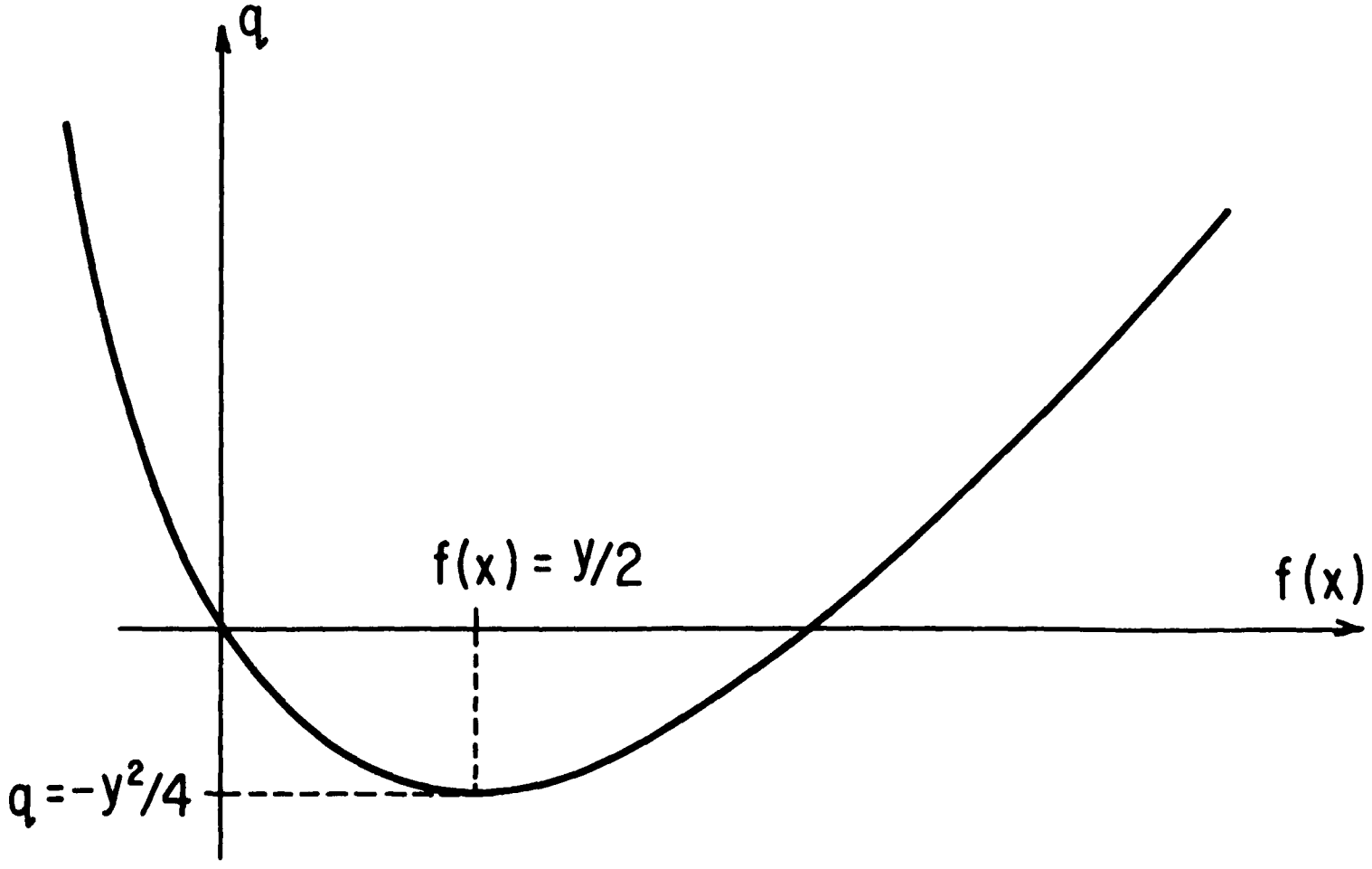


Figure 1. The positivity constraint.

where q_1 is the residual sum of the errors squared between the computed $\vec{a}$ calculated from the solution $\vec{f} = \tilde{\phi}^T\vec{a}$ and the data measurements $\vec{g}$, subject to the constraints of minimizing q_2 (Eq. 19) and of minimizing n quantities $q_3(x_j)$ for the n equations represented by Eq. (21). The inequality sign in Eq. (22) can be replaced by an equality sign, and the problem becomes one of holding constant the quantity q_1

$$q_1 = (\tilde{A}\vec{a} - \vec{g})^T(\tilde{A}\vec{a} - \vec{g}), \tag{23}$$

subject to the constraints of minimizing q_2 and $q_3(j)$. The solution for Eq. (23) and Eqs. (19) and (21) is the vector $\vec{a}$, such that for any i, i = 1,2 ••• p,

$$\frac{\partial}{\partial a_i}\left(q_1 + \gamma_S q_2 + \sum_{j=1}^{n} \gamma_{pj} q_3\right) = 0 \tag{24}$$

where γ_S is the smoothing Lagrange multiplier and γ_{pj} is the j^{th} positivity Lagrange multiplier for f(x) at x_j.

In matrix notation, the solution for $\vec{a}$ (Eq. 24) can be written as:

$$\vec{a} = \left[\tilde{A}^T\tilde{A} + \gamma_S\tilde{H}_S + \sum_{j=1}^{n} \gamma_{pj}\tilde{\phi}\vec{E}_j\vec{E}_j^T\tilde{\phi}^T\right]^{-1}\left[\tilde{A}^T\vec{g} + \sum_{j=1}^{n} \frac{\gamma_{pj}}{2} y_j\tilde{\phi}\vec{E}_j\right]. \tag{25}$$

Equation (25) may be simplified if the positivity constraint is applied equally to each solution point x_j, *i.e.*, $\gamma_{pj} = \gamma_p$ for all j's. Thus, Eq. (25) can be rewritten as:

$$\vec{a} = \left[\tilde{A}^T\tilde{A} + \gamma_S\tilde{H}_S + \sum_{j=1}^{n} \gamma_p\tilde{\phi}\vec{E}_j\vec{E}_j^T\tilde{\phi}^T\right]^{-1}\left[\tilde{A}^T\vec{g} + \gamma_p \sum_{j=1}^{n} \frac{y_j}{2}\tilde{\phi}\vec{E}_j\right]. \tag{26}$$

Let the positivity constraint matrix $\tilde{H}_p$ be:

$$\tilde{H}_p = \sum_{j=1}^{n} \tilde{\phi}\vec{E}_j \vec{E}_j^T \tilde{\phi}^T \tag{27}$$

and the positivity constraint vector $\vec{h}_p(\vec{y})$ be:

$$\vec{h}_p(\vec{y}) = \sum_{j=1}^{n} \frac{y_j}{2} \tilde{\phi}\vec{E}_j, \tag{28}$$

where the $\vec{y}$ column vector (n×1) elements are y_j, j = 1,2 ••• n.

Combining Eqs. (26), (27), and (28) yields the final form of the solution to the basic Eq. (2)

$$\vec{a} = \left[\tilde{A}^T\tilde{A} + \gamma_S\tilde{H}_S + \gamma_p\tilde{H}_p\right]^{-1}\left[\tilde{A}^T\vec{g} + \gamma_p\vec{h}_p(\vec{y})\right] \tag{29}$$

and

$$\vec{f} = \tilde{\phi}^T\vec{a} \quad , \tag{30}$$

where γ_S and γ_p are undetermined Lagrange multipliers.

Equations (29) and (30) are solved by an iterative process. It is assumed initially that any given x in f(x) is equally likely to be present; therefore, a flat size distribution function will be used as a first-guess solution for $\vec{f}$. A first-iteration vector, $\vec{y}^{(1)} = 2\vec{f}$, is substituted in Eq. (29), where a first-iteration vector $\vec{a}^{(1)}$ is calculated. A superscript denotes the iteration number. For the next iteration $\vec{f}^{(2)}$ is calculated from $\vec{a}^{(1)}$ through Eq. (30), and $\vec{y}^{(2)} = 2\vec{f}^{(1)}$ [for positive elements in $\vec{f}^{(1)}$] is substituted in Eq. (29) to solve for a second iterative solution $\vec{a}^{(2)}$.

The process is repeated where the constraint equation, Eq. (20), is used to force the solution $\vec{f}$ towards $\vec{y}/2$ (all elements of $\vec{y}$ are positive) so that any negative values for f(x) that may appear are encouraged to

become positive in successive iterations by the positivity constraint. The iteration process is stopped when all the $\vec{f}$ elements are positive and the computed values of $\vec{g}$ from this solution agree with the measurements for each y_i within some predetermined accuracy.

The significance of negative elements in the solution $\vec{f}$ should be checked by setting negative values to be zero to create a positive vector $\vec{f}$ and then computing two measurement vectors from both solutions. In cases where the two measurement vectors differ by less than the errors, it can be concluded that the negative elements in $\vec{f}$ are insignificant.

The numerical values of the undetermined Lagrange multipliers γ_S and γ_p are chosen for the iterative process as follows. When Eq. (29) is used only with the positivity constraint ($\gamma_S = 0$), γ_p is chosen for the iterative process such that $\gamma_p \tilde{H}_p < \tilde{A}^T\tilde{A}$ and such that the ratio of the largest to the smallest eigenvalues of the matrix $(\tilde{A}^T\tilde{A} + \gamma_p\tilde{H}_p)$ is minimal for γ_p satisfying the first condition. This latter condition decreases the relative error magnification.

If Eq. (29) is used with both Lagrange multipliers, γ_S and γ_p, γ_S and γ_p were chosen such that $\gamma_S\tilde{H}_S \approx \gamma_p\tilde{H}_p$ in order that both constraints will affect $\tilde{A}^T\tilde{A}$ in Eq. (29), and the condition number of $(\tilde{A}^T\tilde{A} + \gamma_S\tilde{H}_S + \gamma_p\tilde{H}_p)$ is minimal for $\gamma_S\tilde{H}_S + \gamma_p\tilde{H}_p < \tilde{A}^T\tilde{A}$.

Finally, the limits [a,b] in Eq. (1) for the inversion process should be chosen with care. Normally they are not known accurately, and their choice can seriously affect the solution vector. The upper limit [b] should normally be decreased if the iteration process cannot produce both a positive $\vec{f}$ and an agreement between the computed measurements and the input measurements.

1.3 Standard Deviation of the Solution

Assuming that there is no error in the mathematical model and in the basis functions, the effect of random noise in the measurements on the standard deviation of the solution points will be estimated.

Following Frieden's[22] derivation for the least-squares problem, Eqs. (29) and (30) are combined to give

$$\hat{\vec{f}} = \tilde{\Phi}^T \left[\tilde{A}^T\tilde{A} + \gamma_S\tilde{H}_S + \gamma_p\tilde{H}_p\right]^{-1} \left[\tilde{A}^T\vec{g} + \gamma_p\vec{h}_p(\vec{y})\right]. \tag{31}$$

Let the matrix $\tilde{D}^{-1}$ be

$$\tilde{D}^{-1} = \tilde{\Phi}^T \left[\tilde{A}^T\tilde{A} + \gamma_S\tilde{H}_S + \gamma_p\tilde{H}_p\right]^{-1} \tag{32}$$

so that Eq. (31) can be written as:

$$\tilde{D}\hat{\vec{f}} = \tilde{A}^T\vec{g} + \gamma_p\vec{h}_p(\vec{y}) \tag{33}$$

where the circumflex indicates an estimation. The error in $\vec{g}$ is modeled as a Gaussian additive noise such that $g_j = N(g_j^S, \sigma_j^2)$, where g_j^S is the exact j^{th} measurment and σ_j^2 is the variance about g_j^S and N represents normal statistics. Taking the expectation of Eq. (33) produces:

$$\tilde{D}\langle\hat{\vec{f}}\rangle = \tilde{A}^T\vec{g}^S + \gamma_p\vec{h}_p(\vec{y}) \tag{33A}$$

Subtracting Eq. (33A) from Eq. (33) yields:

$$\tilde{D}(\hat{\vec{f}} - \langle\hat{\vec{f}}\rangle) = \tilde{A}^T\vec{u} \tag{34}$$

where $\vec{u}$ is the error measurement vector whose elements obey statistics of the form $N(0,\sigma_i^2)$. Decomposing Eq. (34) yields:

$$\sum_j D_{ij}(\hat{f}_j - \langle\hat{f}_j\rangle) = \sum_j A_{ij}^T u_j \tag{35}$$

In a similar way, Eq. (34) yields:

$$\sum_{j'} D_{i',j'}(\hat{f}_{j'} - \langle\hat{f}_{j'}\rangle) = \sum_{j'} A_{i',j'}^T u_{j'} \tag{36}$$

Multiplying Eqs. (35) and (36) and taking the expectation of the product yields (in matrix notation):

$$\tilde{D}\,\tilde{\Gamma}_{\hat{f}}\tilde{D}^{T} = \tilde{A}^{T}\tilde{\Lambda}\tilde{A} \tag{37}$$

where $\tilde{\Gamma}_{\hat{f}}$ is the covariance matrix of the expected solution $\hat{\hat{f}}$ of which the elements are $(\tilde{\Gamma}_{\hat{f}})_{ij} = \langle(\langle\hat{f}_i\rangle - \hat{f}_i)(\langle\hat{f}_j\rangle - \hat{f}_j)\rangle$ and $\tilde{\Lambda}$ is a diagonal matrix of which the elements are the expected values of the variances of the measurement errors σ_i^2. Solving for the covariance matrix $\tilde{\Gamma}_{\hat{f}}$ yields:

$$\tilde{\Gamma}_{\hat{f}} = \tilde{D}^{-1}\,\tilde{A}^{T}\,\tilde{\Lambda}\,\tilde{A}^{T}\,(\tilde{D}^{-1})^{T} \tag{38}$$

where the identity $(\tilde{D}^{-1})^{T} = (\tilde{D}^{T})^{-1}$ was used.

The matrix $\tilde{\Gamma}_{\hat{f}}$ may be determined from Eq. (38) with the matrix $\tilde{D}^{-1}$ determined from Eq. (32). The square root of the diagonal elements of the matrix $\tilde{\Gamma}_{\hat{f}}$ are the standard deviation of the solution points.

The standard deviations of the expected solutions $\hat{\hat{f}}$ are a result of the mapping of error from the measurement space onto the solution space. This mapping depends on the numerical value of the Lagrange multipliers γ_S and γ_p (through the matrix $\tilde{D}^{-1}$), as well as the matrix $\tilde{A}^{T}\tilde{A}$.

Qualitatively, it can be seen that, as the numerical values of γ_S and γ_p increase, elements of the $\tilde{D}^{-1}$ matrix become smaller so that the standard deviations of the solution points decrease. The explanation for this behavior is the fact that the model, including the basis functions and the constraint equations, is assumed to be correct and unbiased. Increasing γ_S and γ_p decreases the condition number of $\tilde{A}^{T}\tilde{A}$ and thereby decreases the error magnification. However, as γ_S and γ_p increase, the measurements computed from the solution will tend to

deviate more from the data measurements. The behavior of the standard deviation of $\hat{\vec{f}}$ suggests that it is better to use as large a numerical value as possible, consistent with the condition stated in Section 1.2 for γ_S and γ_p.

1.4 Basis Functions

The basis functions are crucial to the solution process because the solution [Eq. (30)] is a linear sum of the basis functions and because the numerical quadrature matrix $\tilde{A}$ [Eq. (17)] depends on the form of the basis functions. Assuming a known distribution function n(x), the coefficients a_i for the constructed distribution f(x) [Eq. (13)] can be obtained by using the orthonormal properties of the basis functions

$$a_i = \int_a^b n(x)\phi_i(x)\,dx \tag{39}$$

$$f(x) = \sum_{i=1}^{p} a_i \phi_i(x) \tag{40}$$

Any solution f(x) obtained from an inversion process cannot be closer (in a least-squares sense) to n(x) (the true solution) than the solution constructed from Eqs. (39) and (40). The constraint equations and the kernels can, at most, produce a coefficients set which is the one calculated in Eq. (39), where the coefficients, a_i, are computed from the true solution, n(x).

The solution f(x) can be obtained from a set of basis functions constructed from the kernel functions of the problem by a Gram-Schmidt process.[20] These basis functions will hereafter be referred to as

natural basis functions because they originate directly from the problem. Another possible source for basis functions is empirical basis functions, which will later be referred to as "pseudo"-empirical basis functions.

When using natural basis functions in addition to a smoothing constraint, the solution f(x) will be the smoothest solution consistent with the measurements which may be constructed from the basis functions. When empirical basis functions are used, the solution f(x) lies partially outside the function space of the measurements (kernels). Therefore, the use of empirical basis functions may act as additional information to the measurements. Conversely, the use of empirical basis functions may eliminate many possible mathematical solutions for f(x). However, the constraining effect of empirical basis functions can be minimized if they are chosen in such a way that they lie in the function space of the anticipated distribution functions. Therefore, in order to make use of empirical basis functions, some prior knowledge about the anticipated distributions must be obtained.

1.4.1 Natural Basis Functions

Analysis of the natural basis functions can give an insight into the inherent limitation of a solution obtained from the inversion process solely from the measurements, when no additional information is used. This analysis is important in order to determine what part of the solution f(x) and its computed properties is a result of the inversion process and of the additional information used, and what part is a direct result of the measurements.

If the anticipated solutions are functions n(x), then the type of solutions that can be inferred from measurements calculated from

$g_i(y) = \int k_i(x)n(x)\,dx$ can be computed from Eqs. (39) and (40), when the kernels of the problem are used to compute the basis functions [*i.e.*, the set $\phi_i(x)$]. When an optimal set of kernels is used [*i.e.*, a set of kernel functions which have maximum linear independency between them], the reconstructed functions f(x) show the parts of n(x) to which the measurements are sensitive.

The information about f(x), such as different moments of f(x) (*e.g.*, total number of particles, mean radius, total surface area, and total volume, when f(x) is an aerosol size distribution) can be obtained by an operation of a weighting function w(x) on f(x), *i.e.*, $\int_a^b w(x)f(x)\,dx$. The weighting functions w(x) for the quantities listed above are, respectively, $1, x, x^2, x^3$. $w(x) = x$ will result in a mean radius if $\int_a^b f(x)dx = 1$. The process of deducing the various moments of f(x) from measurements can be viewed as an *approximation* of the appropriate weighting function w(x) by a linear combination of the kernels for which the measurements have been made.[23] Thus, for a general weighting function w(x):

$$w(x) \cong \sum_{i=1}^{m} a_i k_i(x) \tag{41}$$

and

$$\int_a^b w(x)f(x)\,dx \cong \sum_{i=1}^{m} a_i g_i \ . \tag{42}$$

The degree of approximation for w(x) in Eq. (41) will be defined as:

$$q = \left[\frac{\int_a^b e^2(x)\,dx}{\int_a^b w^2(x)\,dx} \right]^{1/2} , \tag{43}$$

where $e(x) = w(x) - w'(x)$ and $w'(x)$ is the approximation to $w(x)$ using Eq. (41). If the approximation of $w(x)$ results in a big error $e(x)$, the corresponding moment of $f(x)$ and the physical quantity it describes cannot be recovered with much reliability.

An estimate of the error in deducing average quantities such as average number density $f(x)$, average x^2, and average x^3 in intervals Δx can be calculated if an appropriate $w(x)$ is chosen. For an average quantity x^n between x_1 and x_2, $w(x)$ will be

$$w(x) = \begin{cases} \dfrac{x^n}{\int_{x_1}^{x_2} x^n \, dx} & x_1 \leq x \leq x_2 \\ 0, & \text{elsewhere.} \end{cases} \tag{44}$$

1.4.2 Empirical Basis Functions

In cases where the use of natural basis functions produces a less than satisfactory result, it is appropriate to search for a set of basis functions that will produce better results.

Given a set of functions $f_i(x)$, $i = 1,2 \cdots N$, a set of basis functions $\phi_i(x)$, $i = 1,2 \cdots N$ can be constructed such that any of the $f_i(x)$ functions can be obtained by using a linear combination of the basis function set such as:

$$f_i(x) = \sum_{j=1}^{N} b_{ij} \phi_j(x) \, . \tag{45}$$

The basis functions $\phi_j(x)$ can be constructed from the eigenvalues and

$g_i(y) = \int k_i(x)n(x)\,dx$ can be computed from Eqs. (39) and (40), when the kernels of the problem are used to compute the basis functions [*i.e.*, the set $\phi_i(x)$]. When an optimal set of kernels is used [*i.e.*, a set of kernel functions which have maximum linear independency between them], the reconstructed functions f(x) show the parts of n(x) to which the measurements are sensitive.

The information about f(x), such as different moments of f(x) (*e.g.*, total number of particles, mean radius, total surface area, and total volume, when f(x) is an aerosol size distribution) can be obtained by an operation of a weighting function w(x) on f(x), *i.e.*, $\int_a^b w(x)f(x)\,dx$. The weighting functions w(x) for the quantities listed above are, respectively, $1, x, x^2, x^3$. w(x) = x will result in a mean radius if $\int_a^b f(x)dx = 1$. The process of deducing the various moments of f(x) from measurements can be viewed as an *approximation* of the appropriate weighting function w(x) by a linear combination of the kernels for which the measurements have been made.[23] Thus, for a general weighting function w(x):

$$w(x) \cong \sum_{i=1}^{m} a_i k_i(x) \tag{41}$$

and

$$\int_a^b w(x)f(x)\,dx \cong \sum_{i=1}^{m} a_i g_i \ . \tag{42}$$

The degree of approximation for w(x) in Eq. (41) will be defined as:

$$q = \left[\frac{\int_a^b e^2(x)\,dx}{\int_a^b w^2(x)\,dx} \right]^{1/2} , \tag{43}$$

where $e(x) = w(x) - w'(x)$ and $w'(x)$ is the approximation to $w(x)$ using Eq. (41). If the approximation of $w(x)$ results in a big error $e(x)$, the corresponding moment of $f(x)$ and the physical quantity it describes cannot be recovered with much reliability.

An estimate of the error in deducing average quantities such as average number density $f(x)$, average x^2, and average x^3 in intervals Δx can be calculated if an appropriate $w(x)$ is chosen. For an average quantity x^n between x_1 and x_2, $w(x)$ will be

$$w(x) = \begin{cases} \dfrac{x^n}{\int_{x_1}^{x_2} x^n \, dx} & x_1 \leq x \leq x_2 \\ 0, & \text{elsewhere.} \end{cases} \tag{44}$$

1.4.2 Empirical Basis Functions

In cases where the use of natural basis functions produces a less than satisfactory result, it is appropriate to search for a set of basis functions that will produce better results.

Given a set of functions $f_i(x)$, $i = 1,2 \cdots N$, a set of basis functions $\phi_i(x)$, $i = 1,2 \cdots N$ can be constructed such that any of the $f_i(x)$ functions can be obtained by using a linear combination of the basis function set such as:

$$f_i(x) = \sum_{j=1}^{N} b_{ij} \phi_j(x) \, . \tag{45}$$

The basis functions $\phi_j(x)$ can be constructed from the eigenvalues and

eigenvectors of the covariance matrix $\tilde{C}$ (N×N) (Twomey,[1] pp. 139-143) formed from the original set of functions $f_i(x)$, such that

$$C_{ij} = \int_a^b f_i(x) f_j(x)\, dx \ . \tag{46}$$

The orthonormal basis functions from the set $f_i(x)$ are

$$\phi_i(x) = \lambda_i^{-1/2}\, \vec{u}_i^{T}\, \vec{F}(x) \ , \tag{47}$$

where λ_i and $\vec{u}_i$ are the i^{th} eigenvalue and eigenvector, respectively, of the matrix $\tilde{C}$, and $\vec{F}(x)$ is a column vector (N×1), the elements of which are continuous functions $f_i(x)$, $i = 1,2 \cdots N$. If p basis functions, where $p \leq N$, are used to represent the N functions f(x) in Eq. (45), then the overall fraction of the N functions f(x) accounted for by the p basis functions is

$$\sum_{j=1}^{p} \lambda_j \Big/ \sum_{j=1}^{N} \lambda_j \ . \tag{48}$$

In principle, there are two sources for a library of functions f(x). The first source might be from many measurements of f(x) that were collected over the years, but this type of library is not always available. The second source for a library model of f(x) functions can be obtained by simulating many functions f(x) according to theoretical models.

The set of basis functions for the expansion of f(x) in Eq. (30) is not unique. For example, any two orthogonal unit vectors rotated in an arbitrary angle to the x axis can describe a vector in the x-y plane. Similarly, but in the function space, the nonuniqueness is also true for the orthogonal empirical basis functions.

The criteria for checking a proposed empirical basis function set are: first, that they should be able to construct as many types of the desired functions f(x) as possible; second, that their orientation will be such that the greatest amount of the m-dimensional function space of the kernels $k_i(x)$, i = 1,2 ••• m will be within the p-dimensional function space of the basis functions $\phi_i(x)$, i = 1,2 ••• p; and third, that the basis functions should be able to construct as closely as possible a chosen delta Dirac function. A minimal spread in the construction f(x) should result from using basis functions, and any uncertainty about the location x of the solution number density f(x) should be minimal.

Because the required properties of the basis functions are known, it is possible to construct a library source from mathematical functions f(x) that do not describe existing distributions f(x) but can yield the desired basis functions. If the resulting basis functions fulfill the criteria stated above, these basis functions can be used for the solution of Eq. (1). The resulting basis functions from the mathematical function library are called the "pseudo"-empirical orthogonal functions.

For the case where f(x) is an aerosol size distribution, the library of the source functions f(x) was chosen to include normal distributions, spaced uniformly every 0.15 μm between integration limits [a,b] and with a standard deviation of 0.2 μm, and few Junge-type distributions $f(x) = x^{-(\nu+1)}$, where the ν values are 1, 2, 3, 4, and 5. The computed basis functions were able to simulate many anticipated aerosol size distributions with a very good accuracy.[24]

The orientation of the empirical basis functions relative to the kernel functions determines the numerical quadrature matrix $\tilde{A}$. The components of the basis functions which are nonorthogonal to the kernels will determine to what extent measurements computed directly from f(x) [Eq. (1)] will agree with measurements computed from the coefficient vector $\vec{a}$ and the matrix $\tilde{A}$ [Eq. (16)]. The fraction of the basis functions which is orthogonal to the kernels determines the amount of information about the unknown f(x) distribution which cannot be obtained through the measurements.

The portion of the kernel $k_i(x)$ for measurement i which is within the function space spanned by the p basis functions can be calculated by Eq. (49).

$$q_i = \left[\frac{\sum_{j=1}^{p} \left[\int_a^b k_i(x)\phi_j(x)\, dx \right]^2}{\int_a^b k_i^2(x)\, dx} \right]^{1/2} \qquad i = 1,2 \cdots m . \qquad (49)$$

The portion of the basis function j which is within the function space spanned by the kernel functions can be calculated by

$$q_j = \left[\frac{\sum_{i=1}^{m} \left[\int_a^b \rho_i(x)\phi_j(x)\, dx \right]^2}{\int_a^b \phi_j^2(x)\, dx} \right]^{1/2} \qquad j = 1,2 \cdots p . \qquad (50)$$

where $\rho_i(x)$, i = 1,2 ••• m, are orthonormal basis functions constructed from the kernel functions by a Gram-Schmidt process.

The approximation of $f(x_0)$ as a linear sum of the basis functions results in some uncertainty as to the location of the point x_0 in the

$f(x_0)$. This uncertainty can be written as

$$f(x_0 \pm \Delta x_0) = \sum_{i=1}^{p} a_i \phi_i(x_0) \ , \tag{51}$$

where Δx_0 is the uncertainty in the location x_0 for $f(x)$ at x_0. The magnitude of the uncertainty Δx_0 can be estimated by substituting a very narrow function at x_0 for $n(x)$ in Eq. (39) and constructing the approximated $f(x_0)$ [Eq.(51)]. The limit of a very narrow function is a delta function. In numerical form, the delta function may be approximated by a normal distribution with a very small standard deviation.

Using an empirical basis function is equivalent to changing the kernels $k_i(x)$ to empirical kernels $B_i(x)$. This can be seen when the equations $\vec{g} = \tilde{A}\vec{a}$ [Eqs. (16), (17)] and $\vec{a} = (\tilde{\phi}\ \tilde{\phi}^T)^{-1}\ \tilde{\phi}\vec{f}$ [obtained from Eq. (14)] are combined to yield:

$$\vec{g} = \tilde{A}(\tilde{\phi}\ \tilde{\phi}^T)^{-1}\ \tilde{\phi}\vec{f} \tag{52}$$

or

$$\vec{g} = \tilde{B}\vec{f} \tag{53}$$

where $\tilde{B} = \tilde{A}(\tilde{\phi}\ \tilde{\phi}^T)^{-1}\ \tilde{\phi}$. If the $\tilde{B}$ matrix is computed for very fine x_i intervals, the trapezoidal rule for integration is a good approximation, and Eq. (52) can be assumed to be nearly equal to the integral form, given by

$$g_i = \int_a^b B_i(x) f(x)\ dr \tag{54}$$

where $B_i(x)$ are the rows of $\tilde{B}$.

The unknown solution $f(x)$ is a linear combination of basis functions constructed from the empirical kernel $B_i(x)$. As a result, similar analysis of the properties of the retrieved solution can be performed on $B_i(x)$, as was described for $k_i(x)$.

1.5 Summary

A method for solving the inverse problem was derived. The method uses a library of functions from which a set of orthogonal basis functions is computed. The source of the library can be from a set of observations or a set of mathematical functions, in which case the basis functions are pseudo-empirical orthogonal functions. It is assumed that any unknown solution f(x) may be constructed from a linear sum of these functions. The problem then becomes one of solving for the unknown coefficients of the basis functions. A solution with a smoothing constraint and/or a positivity constraint can be obtained. A solution with the positivity constraint alone can be useful when the unknown is known to be a narrow function or an unsmooth function. Analysis of the information contained in the measurements and the effect of using additional information is given. This type of analysis is important in order to be able to use the solution properly.

1.6 Literature Cited

1. S. Twomey, *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements* (Elsevier, New York, 1977), 243 p.

2. D. Deirmendjian, "A Survey of Light Scattering Techniques Used in the Remote Monitoring of Atmospheric Aerosols," in *Reviews of Geophysics and Space Physics, Vol. 18* (American Geophysical Union, Washington, D.C., 1980), pp. 341-360.

3. J. R. Bottiger, "Intercomparison of Some Inversion Methods on Systems of Spherical Particles," in *Advances in Remote Sensing Retrieval Methods*, A. Deepak, H. Fleming and M. Chahine, Eds. (A. Deepak, Hampton, Virginia, 1985), pp. 581-593.

4. H. E. Fleming, "Comparison of Linear Inversion Methods by Examination of the Duality Between Iterative and Inverse Matrix Methods," in *Inversion Methods in Atmospheric Remote Sensing*, A. Deepak, Ed. (Academic, New York, 1977), pp. 325-355.

5. G. Strang, *Linear Algebra and Its Applications* (Academic, New York, 1980, 2nd Ed.), 414 p.

6. J. H. Wilkinson, *The Algebraic Eigenvalue Problem* (Clarendon, Oxford, 1965), 662 p.

7. C. D. Rodgers, "Retrieval of Atmosphere Temperature and Composition from Remote Measurements of Thermal Radiation," in *Reviews of Geophysics and Space Physics, Vol. 14* (American Geophysical Union, Washington, D.C., 1976), pp. 609-624.

8. W. L. Smith, "Iterative Solution of the Radiation Transfer Equation for the Temperature and Absorbing Gas Profile of an Atmosphere," Appl. Opt. **9**, 1993 (1970).

9. M. T. Chahine, "Determination of the Temperature Profile in an Atmosphere From Its Outgoing Radiance," J. Opt. Soc. Amer. **58**, 1634 (1968).

10. M. T. Chahine, "Inverse Problems in Radiative Transfer: A Determination of Atmospheric Parameters," J. Atmos. Sci. **27**, 960 (1970).

11. S. Twomey, "On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of the Linear System Produced by Quadrature," J. Assoc. Comput. Mach. **10**, 97 (1963).

12. S. Twomey, "The Application of Numerical Filtering to the Solution of Integral Equations Encountered in Indirect Sensing Measurements," J. Franklin Inst. **279**, 95 (1965).

13. H. E. Fleming and D. Q. Wark, "A Numerical Method for Determining Relative Spectral Response of the Vidicons in a Nimbus Satellite System," Appl. Opt. **4**, 337 (1965).

14. D. Q. Wark and H. E. Fleming, "Indirect Measurements of Atmospheric Temperature Profiles from Satellites: I. Introduction," Mon. Wea. Rev. **94**, 351 (1966).

15. B. M. Herman, S. R. Browning, and J. A. Reagan, "Determination of Aerosol Size Distribution from Lidar Measurements," J. Atmos. Sci. **28**, 763 (1971).

16. M. Grimmer, "The Space-Filtering of Monthly Surface Anomaly Data in Terms of Pattern, Using Empirical Orthogonal Functions," Quart. J. Roy. Meteor. Soc. **89**, 395 (1963).

17. J. C. Alishouse, L. J. Crone, H. E. Fleming, F. L. Van Cleef, and D. Q. Wark, "A Discussion of Empirical Orthogonal Functions and Their Application to Vertical Temperature Profiles," Tellus **19**, 477 (1967).

18. C. L. Mateer, "On the Information Content of Umkehr Observations," J. Atmos. Sci. **22**, 370 (1965).

19. W. D. Sellers and D. N. Yarger, "The Statistical Prediction of the Vertical Ozone Distribution," J. Appl. Meteorol. **8**, 357 (1969).

20. R. Courant and D. Hilbert, *Methods of Mathematical Physics, Vol. 1* (Interscience, New York, 1953), 561 p.

21. C. Lanczos, *Applied Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1956), 539 p.

22. B. R. Frieden, *Probability, Statistical Optics, and Data Testing* (Springer-Verlag, Berlin, 1983), Chapter 14.

23. S. Twomey and H. B. Howell, "Some Aspects of the Optical Estimation of Microstructures in Fog and Cloud," Appl. Opt. **6**, 2125 (1967).

24. A. Ben-David and B. M. Herman, "The Inverse Problem and the Pseudo-Empirical Orthogonal Function Method of Solution. Part II: Application," Appl. Opt., this issue (1987).

2. APPLICATION

2.1 Introduction

Information about the aerosol size distribution is important in many different areas of the atmospheric sciences,[1-3] principally due to their effect on optical phenomena and radiative transfer processes. These effects depend on several factors, such as wavelength of the incident radiation, refractive index of the aerosol material, and size distribution of the aerosols.

Given that *a priori* knowledge exists about the refractive index, with some minimal assumptions about the expected aerosol size distribution, measurements of scattered radiation can be used to obtain information about the aerosol size distribution under conditions of independent scattering (*i.e.*, where no permanent phase relation exists between the radiation scattered by two different particles) and where the scattered radiation that undergoes more than one scattering event is negligible (*i.e.*, an optically thin scattering volume).[4]

In the last decade, many methods for inferring aerosol size distribution from optical remotely sensed measurements were developed. They include spectral extinction measurements,[5-11] aureole and forward scattering measurements,[12-17] combined scattering and extinction measurements,[18,19] angular scattering measurements,[20] and backscattered measurements.[21-23] In all of the aforementioned methods, a wide divergence in the accuracy claimed may be observed. A critical review of some of these methods can be found in Deirmendjian.[24]

The aerosol size distribution inferred from solar extinction and solar aureole measurements represents an average size distribution for

the whole atmospheric depth. The aerosol size distribution obtained from backscattering measurements of a pulsed lidar system is a local property of a scattering volume that can be as small as a few cubic meters at a height z.[25] The inferred aerosol size distribution depends on the assumed refractive indices and radii limits of the aerosols for all methods of solution, as well as the inherent assumption that the particles are spherical in shape. The solution obtained from aureole and extinction measurements is less sensitive to the assumed refractive indices than the one from backscattered measurements. The refractive indices used for aureole and extinction techniques should represent some type of average refractive index for the particulates throughout the vertical extent of the aerosol column.

It has been determined that the backscattered spectral measurements contain more information about the particle size distribution than do extinction, aureole, and angular scattering measurements.[26,4] This study investigates the possibility of inferring aerosol size distributions from simulated backscattered measurements, such as would be obtained by monostatic lidar.[27,25] The results and features of the analyses for the maximum accuracy in the inferred solution (assuming spherical shape and known refractive indices) can set an upper limit for accuracy on any solution inferred from spectral extinction, aureole, and angular scattering measurements. It will be assumed that all wavelengths between 0.3 μm and 10.6 μm (in intervals of 0.1 μm) are available for measurements. The aerosol is taken to be a tropospheric, spherical rural aerosol with radii limits from 0.05 to 10.0 μm, for which the residence time is about a week. The wavelength dependent refractive indices are taken from Kent *et al.*[28]

Inversions of backscattered radiation were obtained by Capps *et al.*,[21] in which the solution was constructed from the basis functions of the backscattering kernel. This method does not use any constraints on the solution, and therefore the solution can be oscillatory and negative, even for the very narrow radii limits that were used (0.001 to 1.3 μm). Zuev and Natts[22] use an iterative technique to determine the refractive indices and the size distribution function from multi-wavelength extinction and backscattering cross-sections which are inferred from monostatic lidar measurements. The accuracy of their method (in the atmospheric boundary layer, $z < 1$ km) is stated to be no greater than twice the error in the measurements (using ruby and neodymium and their second harmonics as laser sources). Ben-David and Herman[23] use an iterative technique where an initial guess is built into the kernel function. By successive iterations, a correction vector is calculated and a solution is constructed subject to a smoothing constraint in the solution.

In this work, the psuedo-empirical orthogonal function method[29] is used for inferring size distribution of spherical aerosols with assumed refractive indices. The method uses empirical basis functions from which the solution is constructed subject to a constraint for non-negative solution points and additionally (optionally) to a smoothing constraint upon the solution. The properties, limitations, and accuracy of the method will be shown, along with examples of inversion results for four data sets. Possible applications for the inferred aerosol size distribution will be discussed.

2.2 The Inverse Problem

In a typical monostatic lidar, a pulsed laser is transmitted in

a narrow beam, and a receiver telescope is co-aligned to collect the radiation scattered in the backward direction. The received lidar response may be described in terms of the lidar equation[25]

$$P_\lambda(R,t) = P_{\lambda 0}(t)\ C'/R^2\ \beta_\lambda(R)\exp\left(-2\int_0^R \sigma_\lambda(r)dr\right) \tag{55}$$

where $P_\lambda(R,t)$ is the power received at time, t, from distance, R, and $P_{\lambda 0}(t)$ is the transmitted power, and C' is the instrumental calibration factor. The term $\beta_\lambda(R)$ is the volume backscattering cross-section, and $\sigma_\lambda(R)$ is the volume extinction coefficient, both at wavelength λ and range R. Assuming the measurements of $P_\lambda(R,T)$ are made at optically thin ranges, R, we may neglect the attenuation term in Eq. (55). Then

$$P_\lambda(R,t) = C\beta_\lambda(R) = C\int_a^b K_\lambda(r,m)f(r)dr = g_\lambda \tag{56}$$

where

$$C = C'/R^2 \times P_{\lambda 0}(t)\ .$$

$K_\lambda(r,m)$ is the particle backscattering cross-section for radius r and, with refractive index, m (to simplify the notation, the dependence of the kernel on the refractive index will be omitted), and f(r) is the number of particles of radius r per unit volume per unit interval in r. The measured backscattered flux $P_\lambda(R,t)$ at wavelength λ will be referred to as g_λ to put the notation into the more usual form. If Eq. (56) is written in numerical form, it becomes

$$\vec{g} = \tilde{A}\,\vec{f} \tag{57}$$

where $\vec{g}$ is an m×1 column vector whose elements are the backscattered flux at at m wavelengths, $\tilde{A}$ is an m×n matrix composed of the particle

backscattering cross-sections for the various wavelengths and radii intervals and also contains any numerical quadrature required in addition to the constant C, and $\vec{f}$ is the unknown n×1 column vector whose elements are the number densities at the n discrete radii. For the remainder of this work, we assume that the measurements, g_λ, are given and examine the feasibility and accuracies obtainable in inverting the measurements to obtain the unknown f(r). In addition, we introduce a new inversion approach through the use of pseudo-empirical orthogonal functions to describe the unknown f(r) and also introduce a positivity constraint which helps insure that the values of f(r) so obtained are not physically unreal negative numbers.

To see the difficulties of solving Eq. (56) or its numerical equivalent, Eq. (57), it should be noted that the measurements g_λ are actually equal to inner dot products in the function space (Hilbert space) between the kernels $k_\lambda(r)$ and the unknown function f(r) in Eq. (56) or equally dot products of the row vectors of the $\tilde{A}$ matrix, $\vec{A}_i(j)$ with the unknown column vector $\vec{f}$ (*i.e.*, $g_i = \sum_j A_{ij} f_j$).

The above geometrical viewpoint of Eqs. (56) and (57) addresses the inverse problem thusly: given m projections, g_λ, of an unknown function (vector) f(r) on some set of m skew functions (vectors) $k_\lambda(r)$ to construct the unknown function (vector) f(r). As as result of the erratic fine structure of the kernel (Fig. 2), a large number of solution points n must be taken so that the integral in Eq. (56) can be evaluated. Hence, there are m equations and n unknowns where n may be larger than m (n is usually on the order of 50, and m is on the order of 10).

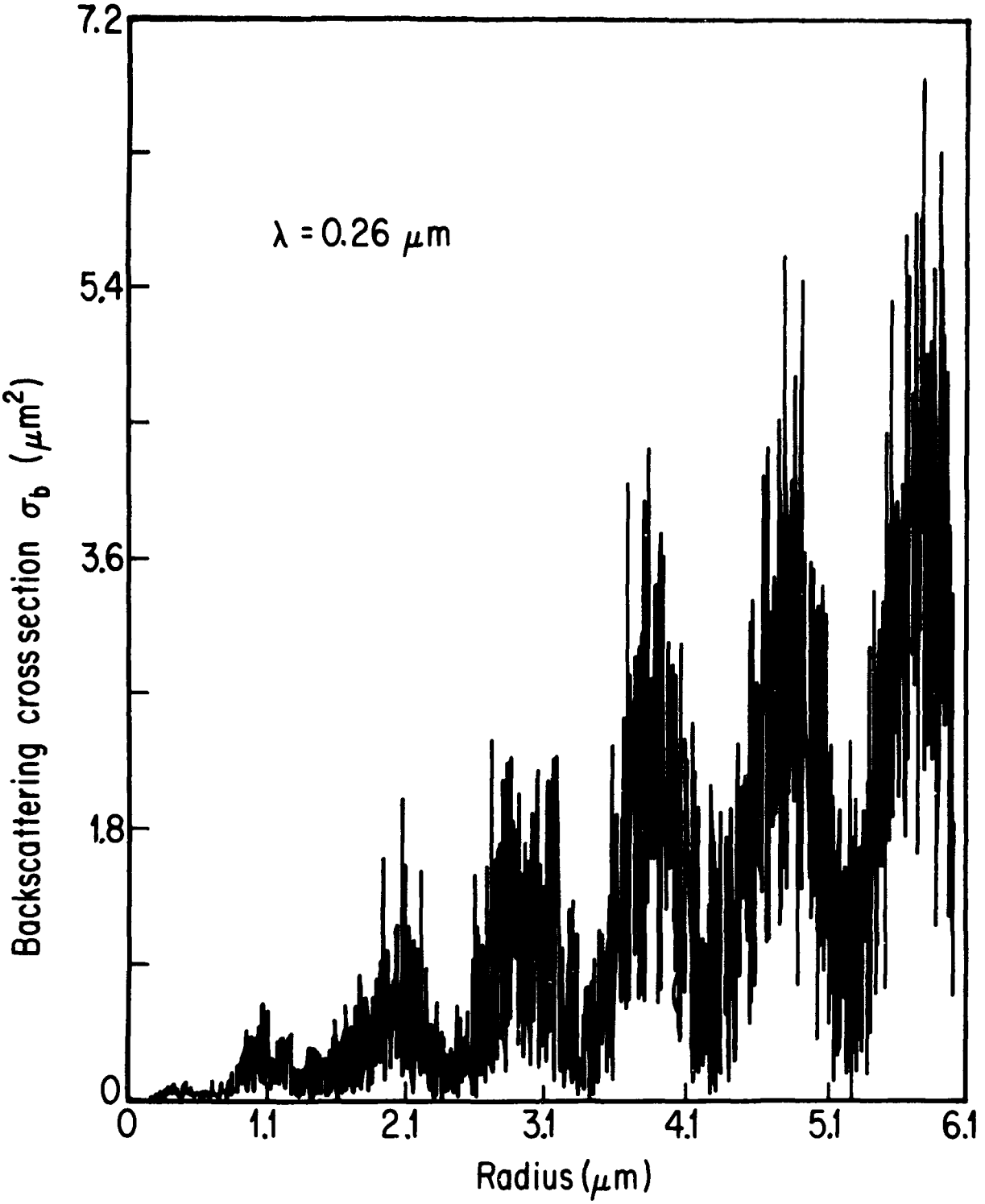


Figure 2. The backscattering cross-section.

2.3 The Information Available in the Measurements

2.3.1 Independence of the Kernel Functions

An analysis of the kernels $k_i(r)$ of the inverse problem can give an insight into the information available in the measurements about the unknown aerosol size distribution. This analysis utilizes results from the previous work by the authors.[29]

Recent developments in laser technology make available a wide range of wavelengths.[25] In order to examine the theoretically maximum information content possible, we assume all wavelengths between 0.3 μm and 10.6 μm in intervals of 0.1 μm may be used.

One hundered and four wavelengths between 0.3 μm to 10.6 μm were thus selected and the backscattering cross-sections computed for each wavelength as a function of size and refractive index of the rural tropospheric aerosol. These functions were used as kernels $k_i(r)$, $i = 1,2,\cdots 104$, and were arranged in order of maximum independency between them. The independency within the kernel functions was measured as the maximum orthogonality between kernel i and all the other kernels $j \neq i$.

The interdependence between the kernels[30,31] $k_i(r)$, $i = 1,2,\cdots m$, can be demonstrated as follows. In principle and from a purely mathematical point of view, two or more of the kernels are linearly dependent if there is a set of coefficients α_i, $i = 1,2,\cdots m$, such that

$$\sum_{i=1}^{m} \alpha_i k_i(r) = 0 \qquad (58)$$

and

$$\sum_{i} \alpha_i^2 = 1$$

in order to eliminate the trivial case.

In a real physical situation, there are uncertainties in the measured quantities or in the mathematical model. In this situation, a linear combination of the kernels that results in the right-hand side of Eq. (58) equalling some value ε, $\varepsilon > 0$ but less than the uncertainties involved, is no better than a zero value and is equivalent to linear dependency between two or more of the kernels.

If it is assumed that the uncertainty in each wavelength i is ε_i, Eq. (58) can be written as:

$$\sum_{i=1}^{m} \alpha_i k_i(r) + \alpha_i \varepsilon_i(r) = q \tag{59}$$

$$\sum_{i=1}^{m} \alpha_i^2 = 1$$

or in vector notation

$$(\vec{k}^T + \vec{\varepsilon}^T)\vec{\alpha} = q$$

$$\vec{\alpha}^T\vec{\alpha} = 1$$

where the $\vec{k}$ vector elements are $k_i(r)$, the $\vec{\alpha}$ vector elements are the coefficients α_i, and the $\vec{\varepsilon}$ vector elements are $\varepsilon_i(r)$ that give the uncertainty in the measurement g_i or represent uncertainties in the physical model (single scattering approximation, refractive index uncertainties, and so forth), and the superscript T denotes a transpose operation.

The quantity to be minimized is

$$|q|^2 = \vec{\alpha}^T(\vec{k} + \vec{\varepsilon})(\vec{k}^T + \vec{\varepsilon}^T)\vec{\alpha} \tag{60}$$

subject to the constraint $\vec{\alpha}^T\vec{\alpha} = 1$.

If the expected value of $\langle|q|\rangle$ (where $\langle|\cdot|\rangle$ denotes an averaging process) is taken and a Gaussian additive noise for ε_i is assumed, such that

$$\varepsilon_i = N[0,\ P_i^2 k_i^2(r)]\ ,$$

where N represents a normal statistic about zero mean and a variance $P_i^2 k_i^2(r)$, P_i is the fraction error in $k_i(r)$ and $\langle \vec{k} \cdot \vec{\varepsilon} \rangle = 0$, Eq. (60) yields:

$$\langle|q|^2\rangle = \vec{\alpha}^T \vec{k}\vec{k}^T \vec{\alpha} + \vec{\alpha}^T \Lambda \vec{\alpha} \qquad (61)$$

where Λ is a diagonal matrix whose elements are

$$\Lambda_{ii} = \int_a^b P_i^2 k_i^2(r)\ dr.$$

Eq. (61) can be written symbolically as $\langle|q^2|\rangle = q_1 + q_2$. The minimum of q_1 is the smallest eigenvalue, λ_{min}, of the matrix C whose elements are

$$C_{ij} = \int_a^b k_i(r)\ k_j(r)\ dr,$$

$$\int_a^b k_i^2(r)\ dr = 1 \qquad \text{(Twomey}^{32}\text{, p. 189)}$$

and

$$q_2 \text{ is } \sum_{i=1}^{m} \alpha_i^2\ P_i^2.$$

Finally, Eq. (61) yields

$$\langle|q|\rangle^2 = \lambda_{min}\ (\text{of } \tilde{C}) + \sum_{i=1}^{m} \alpha_i^2\ P_i^2.$$

The maximum value of q_2 can be calculated from Schwarz's inequality:

$$\sum_{i=1}^{m} \alpha_i^2 P_i^2 \leq \sum_{i=1}^{m} \alpha_i^2 \sum_{i=1}^{m} P_i^2 = mP^2$$

(if we assume all $P_i = P$). Hence, if $\lambda_{min} > mP^2$, the m kernels are independent. To ensure a signal-to-noise ratio of 10, the condition for independence of m kernels which contain an error of magnitude P will be set at

$$\lambda_{min} = 10\ mP^2. \tag{62}$$

The covariance matrix $\tilde{C}$ was computed for various numbers of wavelengths. The covariance matrix' eigenvalues were analyzed [Eq. (62)] to obtain the number of wavelengths that yield independent measurements with some predetermined measurement accuracy. Results are shown in Table 1.

Table 1. Number of independent wavelengths (measurements) and minimum accuracy needed in the measurements.

Number of Wavelengths	Accuracy [%]
40	0.5
35	0.85
30	0.9
25	2.0
20	2.25
15	2.7
5	4.7

2.3.2 Information of Unknown Size Distribution Contained in the Kernels

Direct measurements of aerosol size distributions show that the aerosol size distribution, n(r), can be approximately described by a log normal distribution[33-35] and a power law, $n(r) = cr^{-\nu}$ size distribution.[36-39]

These distributions are recoverable from the measurements only if they are within the function space spanned by the kernels [*i.e.*, Eq. (56) or Eq. (57)]. To examine the extent to which expected size distributions, n(r), lie within the function space of the kernels, the following procedure was followed. A set of orthonormal basis functions (referred to as "natural" basis functions in the previous work) was constructed from the set of kernels. Expected size distribution functions, n(r), were then constructed from combinations of the basis functions. These reconstructed distribution functions, f(r), were then compared to the original n(r) to determine the degree to which the f(r) lie within the space of the kernels.

Figures 3a-f show results of log-normal functions, n(r), and the constructed functions, f(r), calculated by using 40 natural basis functions. The parameters for the log-normal n(r) (standard variation and mean radius) are shown in the figures.

Examination of these figures shows that, while some size distribution functions, n(r), are reproduced reasonably well by the basis functions (*i.e.*, Fig. 3c,d,e,f), others are very poorly reconstructed (*i.e.*, Figs. 4a,b,c,d, and Fig. 3a). Thus, it is evident that some size distributions, most notably power law types (Figs. 4a-d) and very sharply peaked and narrow log-normal types (Fig. 3a) lie primarily outside of the function space of the kernels (*i.e.*, they possess

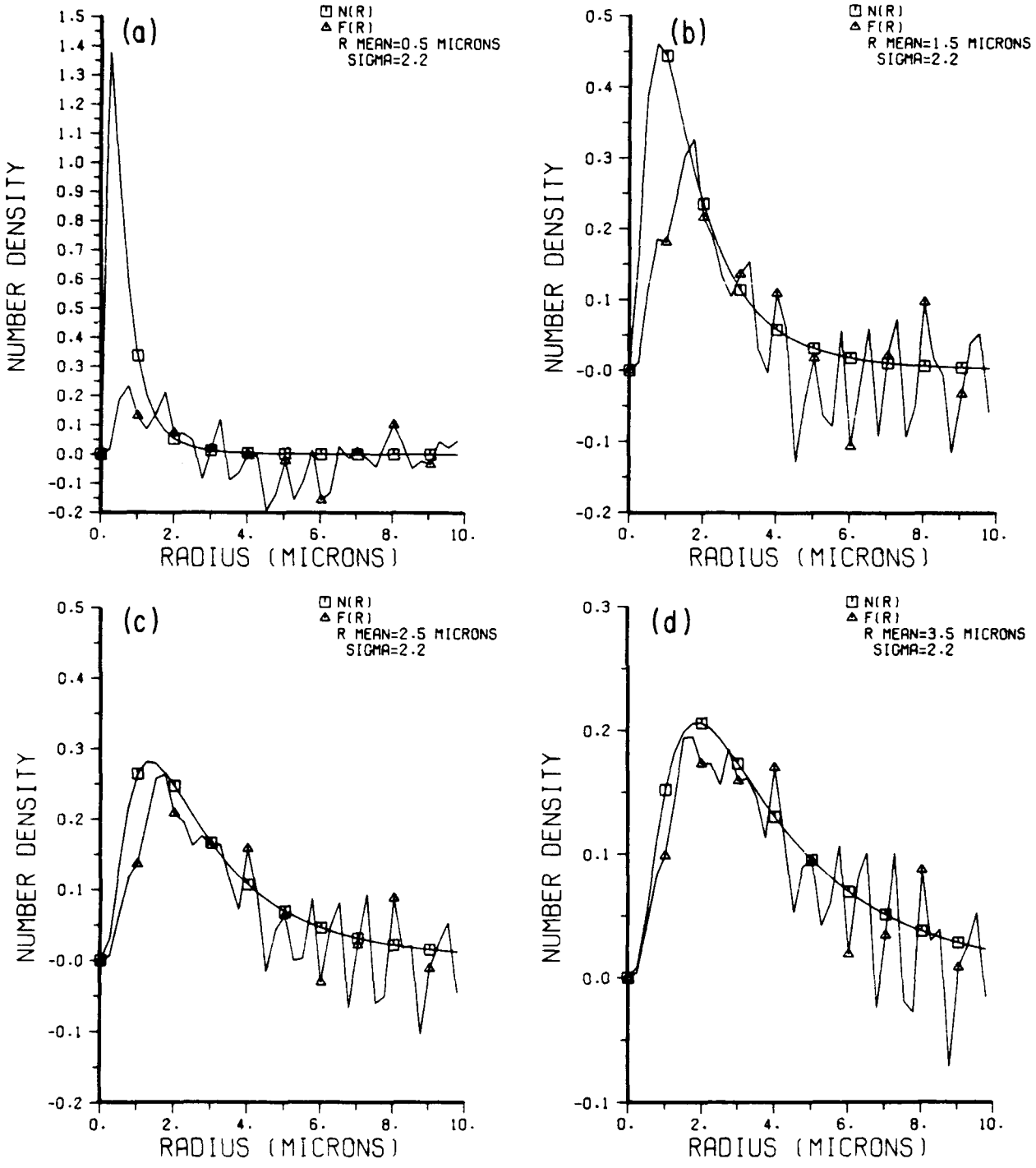


Figure 3. Reconstruction of size distribution f(r) from log-normal size distribution n(r), using 40 natural basis functions.

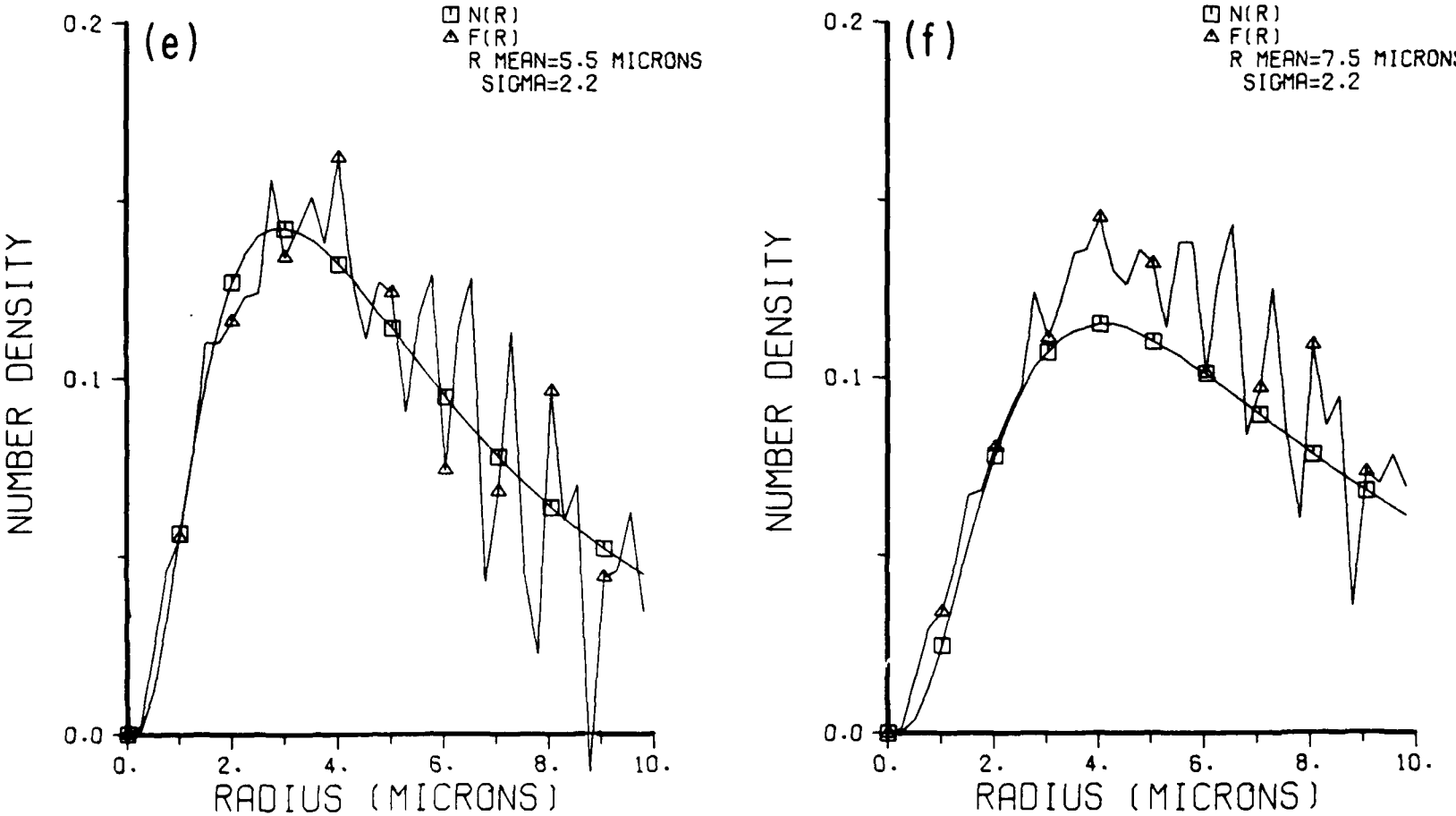


Figure 3 (continued).

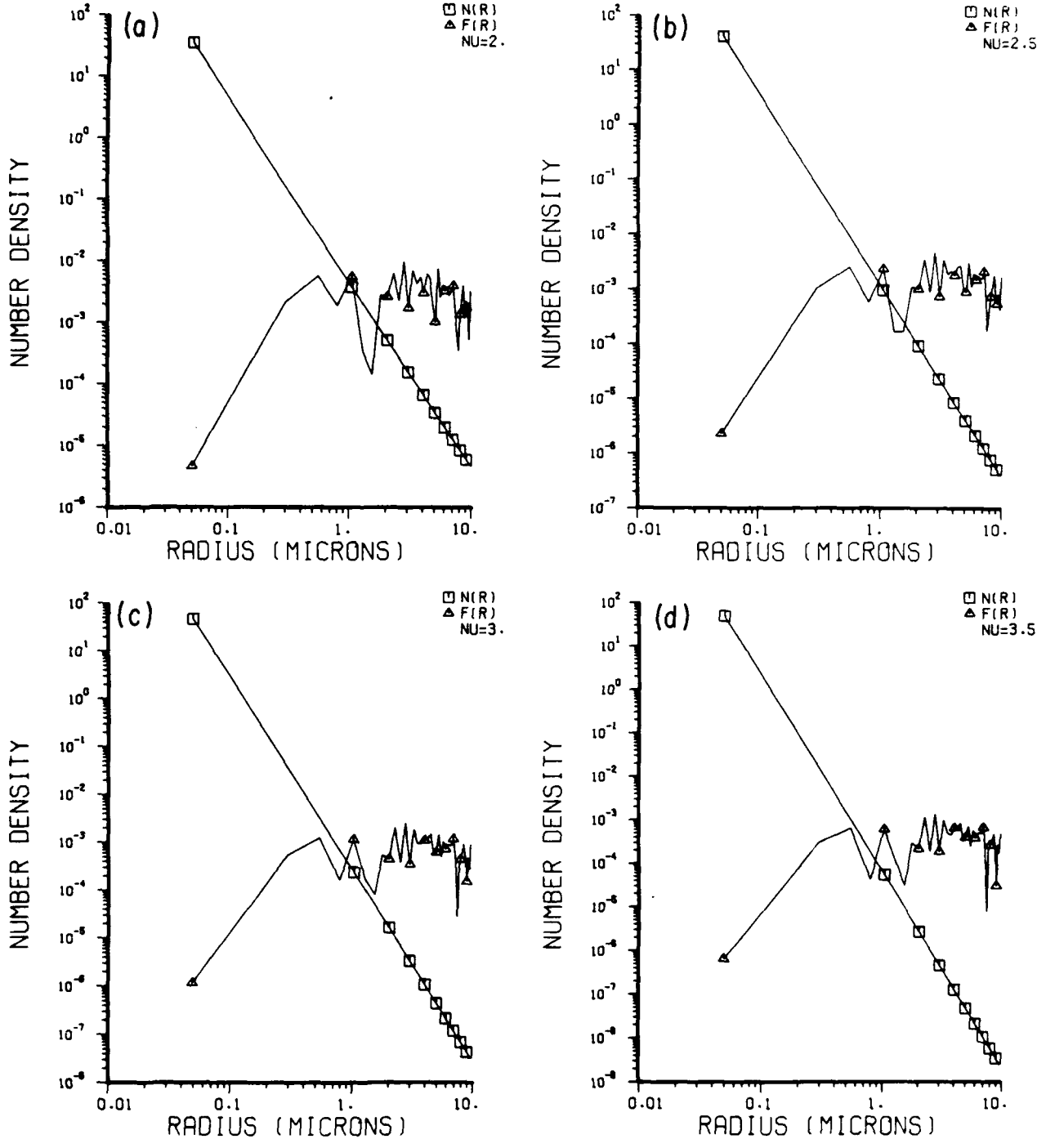


Figure 4. Reconstruction of size distribution f(r) from power-law size distribution n(r), using 40 natural basis functions.

large components which are orthogonal to the kernels) and therefore measurements between 0.3 and 10.6 μm contain limited information as to their form.

An estimate of error in deducing average number density in radii interval Δr = 0.5 μm can be calculated by approximating a weighting function $w(r)$[29, Eq. 44] from the 40 kernel functions. This error was calculated to be 80%.

2.4 Additional Information Contained in the Pseudo-Empirical Orthogonal Functions

A brief review of the method of solution using pseudo-empirical functions[29] will be given before examining the pseudo-empirical orthogonal basis functions used in this work and presenting results of the inversion process.

In the following work, basis functions constructed from a matrix whose elements are composed of a set of mathematical functions (normal and power law functions) are employed, as opposed to the natural basis functions constructed from the kernels as used in the previous section. These basis functions are referred to as pseudo-orthogonal basis functions.

Assuming that the unknown solution $f(r)$ can be constructed from coefficients a_j and basis functions $\phi_j(r)$, Eq. (56) can be written:

$$g_i = \int k_i(r) \sum_j a_j \phi_j(r) dr,$$

or, in matrix notation:

$$\vec{g} = \tilde{A}\vec{a} \tag{63}$$

where $\tilde{A}$ is a matrix whose elements are:

$$A_{ij} = \int k_i(r) \phi_j(r) dr, \tag{64}$$

and $\vec{a}$ is the unknown coefficient vector from which the solution f(r) is constructed, *i.e.*,

$$f(r) = \sum_j a_j \phi_j(r) \tag{65}$$

and $\vec{g}$ is the measurements vector.

A direct solution of Eq. (63) is almost always unstable. For the present method, two types of constraints on the solution are employed in order to solve Eq. (63). The first constraint is a smoothing constraint,[40] such that:

$$\sum_i \frac{\partial^2 f(r_i)}{\partial r_i^2} = \min \ , \tag{66}$$

as the solution is usually expected to be smooth and to filter out artificial oscillations in the solution. The second constraint is a "positivity" constraint, such that $f(r_i) > 0$ for any r_i, as all solution points must physically be positive.

The positivity constraint is employed in an iterative manner. An initial first-guess distribution, y(r), that is positive for all r is used to start the procedure. The expression

$$q = f^{2(1)}(r) - y(r)\, f^{(1)}(r) \tag{67}$$

where $f^{(1)}(r)$ is the first iterative solution, is then minimized as the positivity constraint. For any value of y(r), the minimum value of q is -y(r)/4, for which f(r) = y(r)/2, a positive number. This constraint tends to force the solution toward y/2 for the first iteration. The degree of forcing depends on the strength given to the constraint. For the next iteration, $y^{(1)}(r)$ is set to be equal to 2 f(r), and the process is repeated. Any negative values of f(r) which may appear are encouraged to become positive in successive iterations.

Using the method of Lagrange multipliers, the final solution employing both constraints is given by:

$$\vec{a} = [\tilde{A}^T\tilde{A} + \gamma_s\tilde{H}_s + \gamma_p\tilde{H}_p]^{-1}\ [\tilde{A}^T\vec{g} + \gamma_p\vec{h}_p(\vec{y})] \tag{68}$$

and

$$f(r) = \sum_i a_i \phi_i(r) \tag{69}$$

or

$$\vec{f} = \tilde{\phi}^T\vec{a} \tag{69a}$$

where $\tilde{H}_p$ and $\tilde{H}_s$ are constraint matrices arising from the minimization criteria, $\vec{h}_p(\vec{y})$ is a positivity constraint vector which is a function of $\vec{y}$, γ_s and γ_p are the Lagrange multipliers which determine the strength of the smoothing and positivity constraints, and $\tilde{\phi}$ is a matrix whose rows are $\phi_i(r)$. T denotes a transpose operation, as before.

The standard deviation of the solution is the square root of the diagonal elements of the matrix $\tilde{\Gamma}_{\hat{f}}$, where

$$\tilde{\Gamma}_{\hat{f}} = \tilde{D}^{-1}\ \tilde{A}^T\Lambda A^T(\tilde{D}^{-1})^T \tag{69b}$$

and

$$\tilde{D}^{-1} = \tilde{\phi}^T[\tilde{A}^T\tilde{A} + \gamma_s\tilde{H}_s - \gamma_p\tilde{H}_p]^{-1}$$

and Λ is a diagonal matrix whose elements are the expected values of the variances of the measurements errors.

The basis functions $\phi_j(r)$ can be constructed from the eigenvalues and eigenvectors of the covariance matrix $\tilde{C}$ (N×N) (Twomey,[32] pp. 139-143) formed from the original set of functions $f_i(r)$, such that

$$C_{ij} = \int_a^b f_i(r)\ f_j(r)dr \tag{70}$$

The orthonormal basis functions from the set $f_i(r)$ are given by

$$\phi_i(r) = \lambda_i^{-1/2} \vec{u}_i^T \vec{F}(r) \tag{71}$$

where λ_i and $\vec{u}_i$ are the i^{th} eigenvalue and eigenvector, respectively, of the matrix $\tilde{C}$, and $\vec{F}(r)$ is a column vector (N×1) the elements of which are continuous functions $f_i(r)$, i = 1,2•••N.

Figures 3 and 4 showed that the use of natural basis functions produces a less than satisfactory result. Therefore, it is logical to search for a set of basis functions that will produce better results.

The criteria for checking a proposed empirical basis function set are: first, that they should be able to construct as many types of aerosol size distribution functions as possible; second, that their orientation will be such that the greatest amount of the m-dimensional function space of the kernels $k_i(r)$, i = 1,2•••m will be within the p-dimensional function space of the basis functions $\phi_i(r)$, i = 1,2•••p; and thirdly, the basis functions should be able to closely approximate a chosen delta Dirac function, in order that a minimal spread in the construction f(r) results from using the basis functions, and any uncertainty about the location r of the solution number density f(r) be minimal.

It is standard procedure to compute basis functions from a large library of actual, measured functions. However, in the present case, since an adequate library of measured aerosol size distributions does not exist, we construct a "library" based upon expected forms of the unknown distributions, *i.e.*, the f(r) functions - hence, the name "pseudo"-empirical orthogonal functions.

The library of the source functions f(r) was chosen to include 68 functions (Figure 5). Sixty-three functions were very narrow normal distributions, spaced uniformly every 0.15 μm between 0.2 μm and 9.5 μm and with a standard deviation of 0.2 μm. The remaining 5 functions were chosen to be Junge-type distributions $f(r) = r^{-(\nu+1)}$, where the ν values are 1, 2, 3, 4, and 5. All 68 functions were calculated for a radius range between 0.05 μm and 10 μm. Although the first 30 basis functions can account for 99% of the overall variation of the 68 functions, the errors resulting in the measurements computed from the constructed Junge-type distributions and the measurements computed from the actual Junge distributions were of several orders of magnitude. Most of the information about Junge-type distribution is contained in the last few basis functions. Therefore, all 68 basis functions computed from Eq. (71) are used.

In principle, a larger set of source functions will result in a better quality of basis functions according to the specified properties mentioned above. This source set of equations will work if the normal distribution functions are as narrow as possible and if there is an overlap between the functions so that the resulting basis functions will be continuous functions. However, as the set of normal source functions becomes more numerous, more basis functions are required in order to approximate Junge-type distributions and their measurements. Hence, the dimension of the $\tilde{A}^T\tilde{A}$ matrix will be bigger and the computation time needed for the iterative process will increase considerably. Furthermore, the larger the set of source functions, the smaller the smallest eigenvalues become, resulting in poor accuracy in computing the basis functions in Eq. (71).

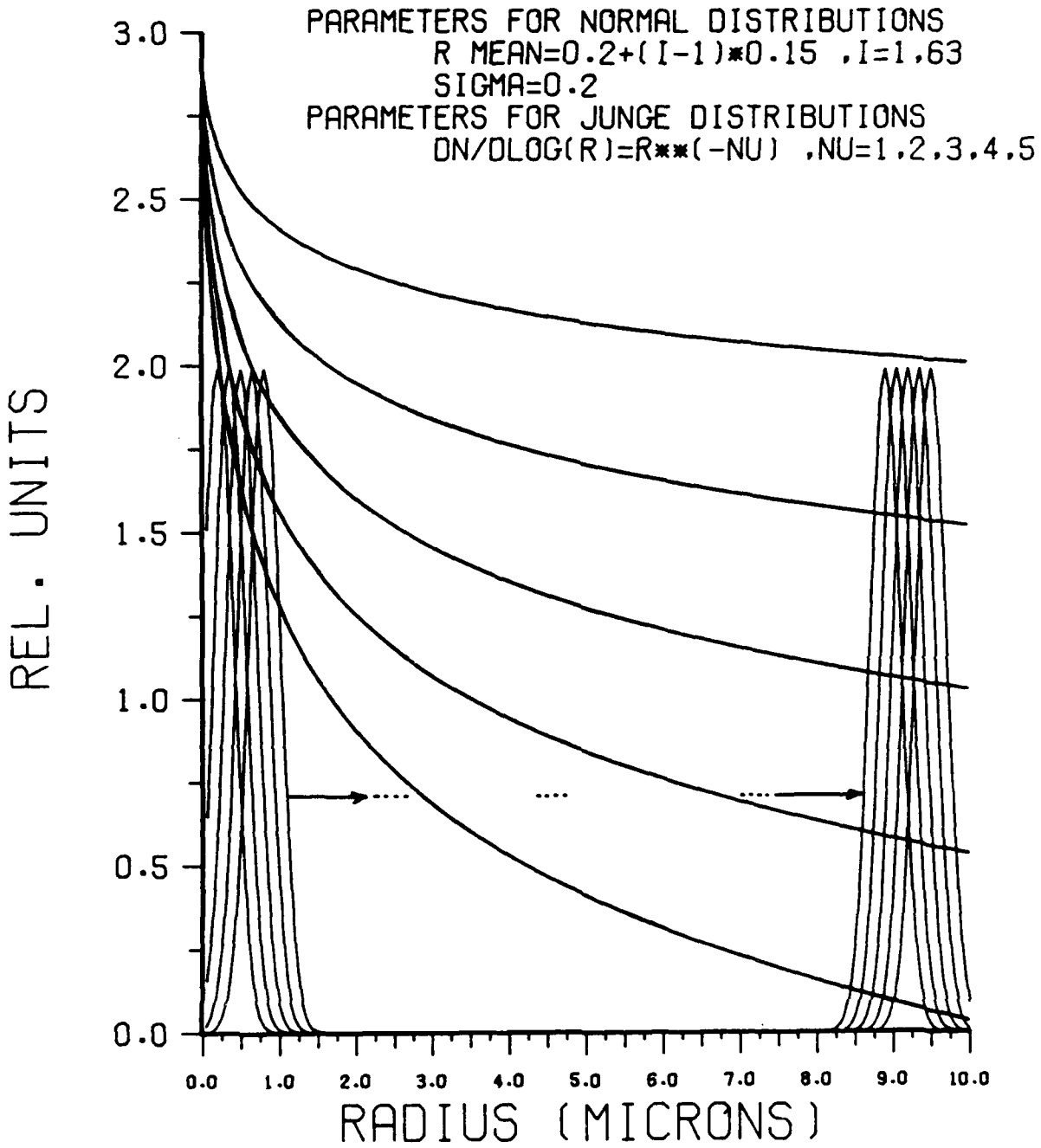


Figure 5. The "pseudo"-empirical functions f(r).

Therefore, an optimal number of functions must be determined which is as large as possible and still allows accurate determination of the smallest eigenvalues. Based on the factors heretofore described, the basis functions chosen represent an optimized set.

Figure 6a-e presents results of reconstructed aerosol size distributions f(r) from various aerosol size distribution models n(r)[41] that differed from the original source functions. The reconstructed distributions were calculated from the pseudo-empirical orthogonal functions from the equation

$$f(r) = \sum_{i=1}^{p} a_i \phi_i(r) \tag{72}$$

where

$$a_i = \int_a^b n(r)\phi_i(r)dr \tag{73}$$

The solid curves are the analytic models, n(r), and the symbols represent the reconstructed f(r). It can be seen that, in most cases, the symbols fall on the solid curves, which is to say that $f(r) \cong n(r)$.

Figure 6a represents reconstruction of various log-normal distributions given by

$$n(r) = 1/(2\pi)^{1/2} \exp[-1/2\ (\ln r - \ln \bar{r})^2/\ln^2\sigma] \tag{74}$$

which is believed to represent the size distribution function for aerosols having soil-derived components.[33] Figure 6b shows reconstructions of various Junge-type distributions given by

$$n(r) = r^{-(\nu+1)} \tag{75}$$

which was proposed by Junge to represent continental aerosol. The

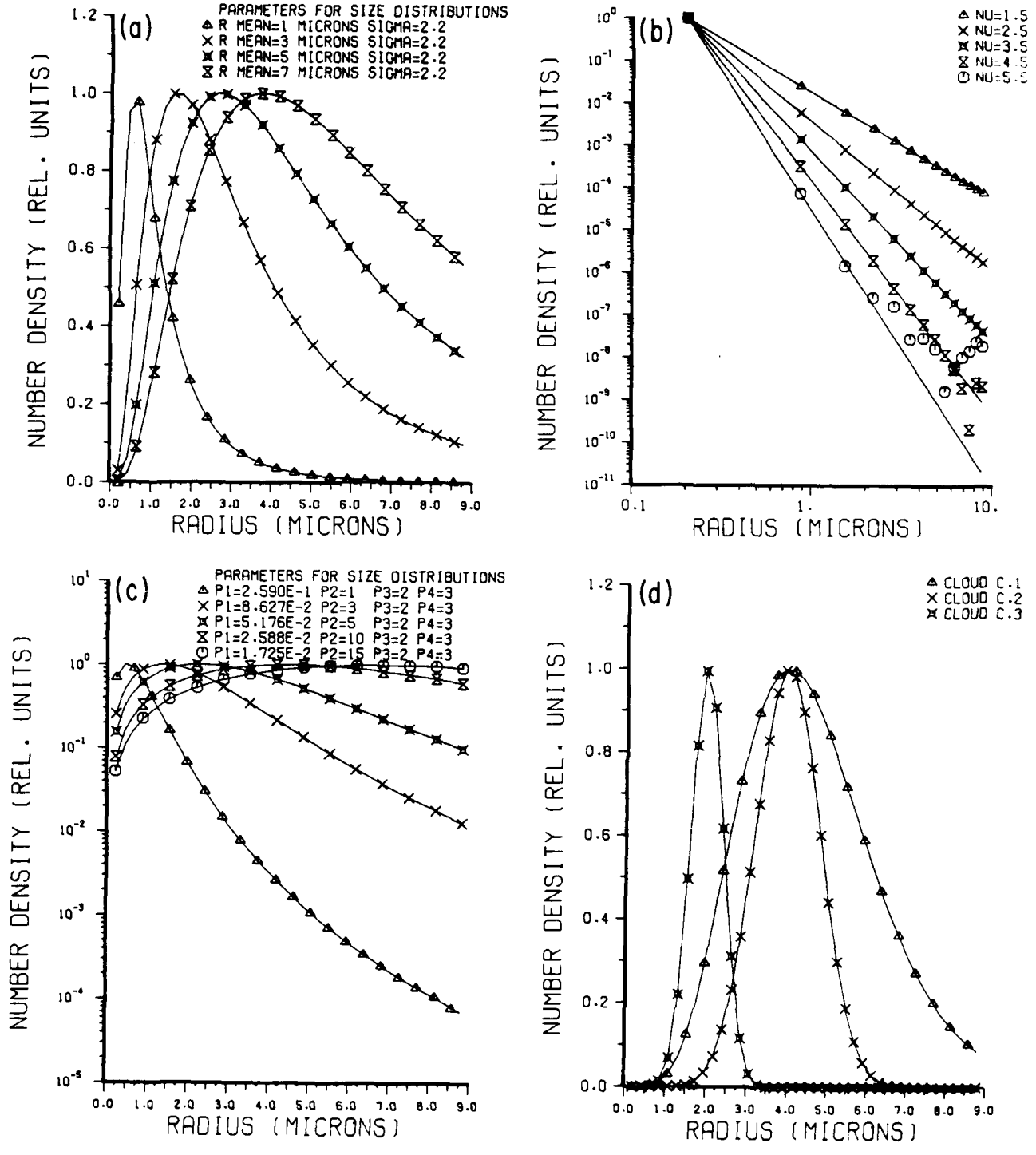


Figure 6. Reconstruction of size distribution f(r) from aerosol size distribution n(r), using the "pseudo"-empirical basis functions (a) for log-normal distribution, (b) for power-law distribution, (c) for regularized power-law distribution, (d) for modified Gamma distribution and (e) for inverse modified Gamma distribution.

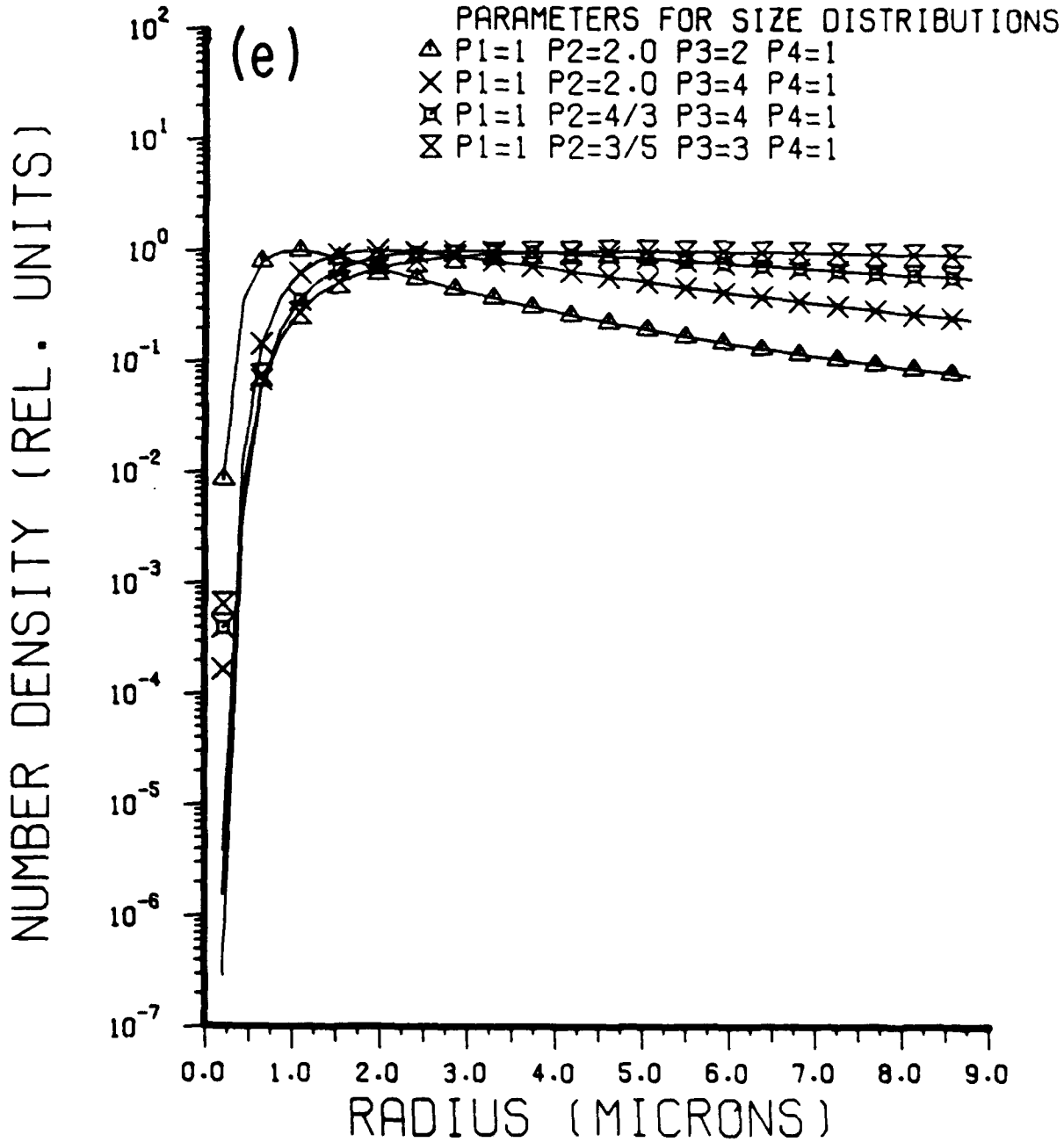


Figure 6 (continued).

numerical values for ν are different from the numerical values in the source function set (Figure 5). An inspection of Figure 6 reveals that when n(r) changes more than 7 orders of magnitude, the constructed distribution f(r) (symbols) deviates from n(r) (solid line). However, measurements computed from the two distributions agree within 5%. In cases where a significant contribution to the measurements is from number densities with a dynamic range of more than seven orders of magnitude, the computed measurements from f(r) will deviate significantly from measurements computed from n(r).

In Figure 6c, n(r) is a regularized power law distribution given by

$$n(r) = p_1/p_2 \, [(r/p_2)^{p_3-1}]/[1 + (r/p_2)^{p_3}]^{p_4} \tag{76}$$

where p_1 through p_4 are constants. This distribution is similar to the Junge-type distribution at large radii, but does not have a singularity at $r = 0$.

In Figure 6d, n(r) is the so-called modified Gamma distribution, given by

$$n(r) = p_1 r^{p_2} \exp(-p_3 r^{p_4}) \tag{77}$$

where the constants p_1 through p_4 determine various models of aerosols such as Haze H, Haze M, Haze L, and cloud C.1 to C.3 used by Deirmendjian.[42] Because the Haze models (L and H) range to more than 40 orders of magnitude for the radii range 0.05-10.0 μm, only the cloud C.1 to C.3 are presented.

In Figure 6e, n(r) is the so-called inverse modified Gamma distribution, given by

$$n(r) = p_1 \exp(-p_3/r^{p_4})/r^{p_2} \tag{78}$$

which describes exponential fall-off at small radii and power law behavior at large radii. This type of distribution can represent dry aerosols.

The fraction of the basis functions which is orthogonal to the backscattering kernels determines the amount of information about the unknown aerosol size distribution which cannot be obtained through the measurements.

The portion of the kernel $k_i(r)$ for wavelength λ_i which is within the function space spanned by the p basis functions can be calculated by Eq. (79):

$$q_i = \left[\frac{\sum_{j=1}^{p} \left[\int_a^b k_i(r)\phi_j(r)\, dr \right]^2}{\int_a^b k_i^2(r)\, dr} \right]^{1/2} \qquad i = 1,2 \cdots m . \qquad (79)$$

The portion of the basis function j which is within the function space spanned by the backscattering kernel functions can be calculated by Eq. (80):

$$q_j = \left[\frac{\sum_{i=1}^{m} \left[\int_a^b \rho_i(r)\phi_j(r)\, dr \right]^2}{\int_a^b \phi_j^2(r)\, dr} \right]^{1/2} \qquad j = 1,2 \cdots p . \qquad (80)$$

where $\rho_i(r)$ = 1.2•••m are orthonormal basis functions constructed from the kernel functions.

Figure 7a shows the portion of $k_i(r)$ which is within the function space of the "pseudo"-empirical basis functions. It can be seen that,

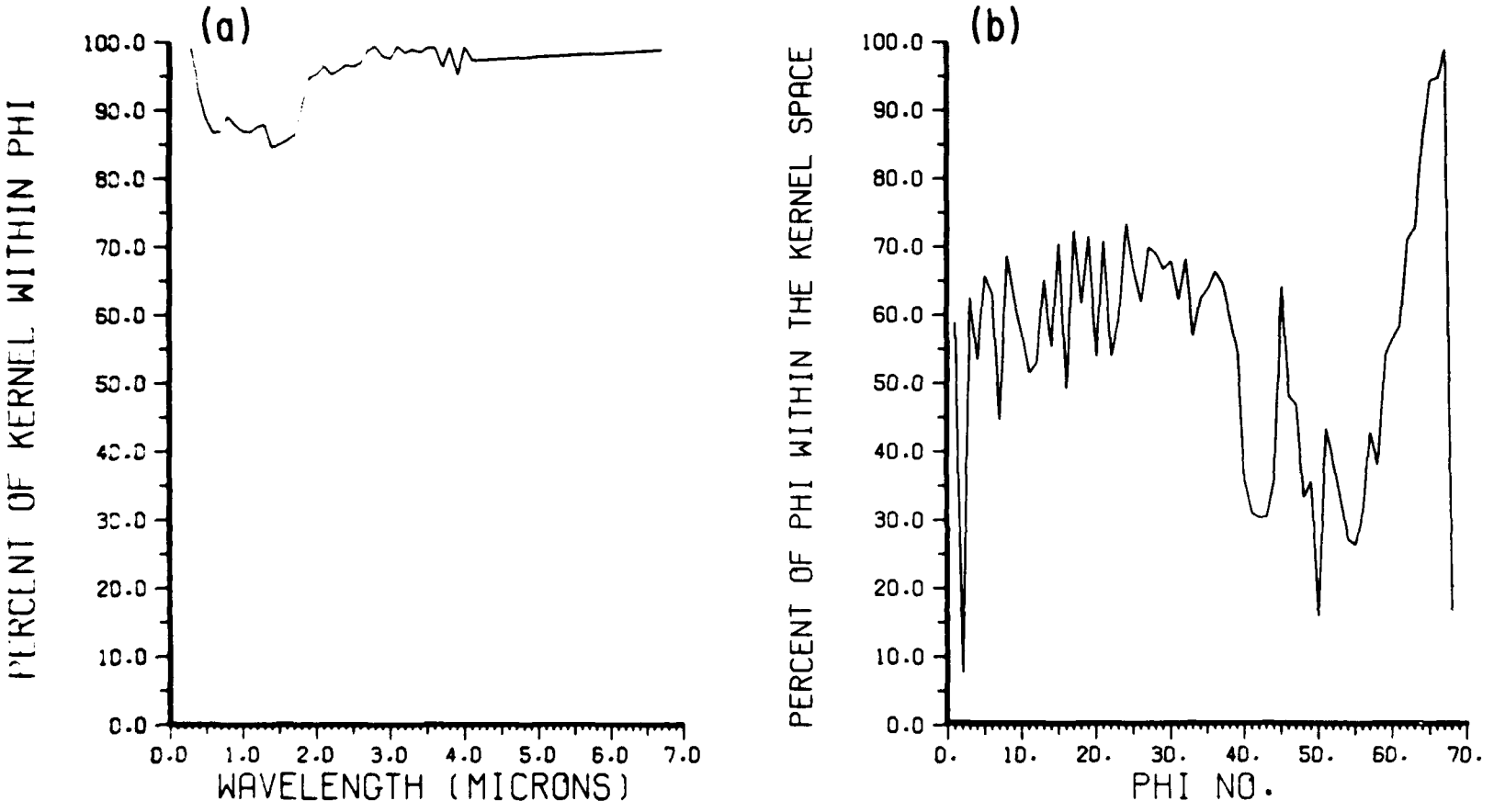


Figure 7. The orientation of the "pseudo"-empirical basis functions relative to the kernels. (a) Portion of K_λ within the empirical basis functions, and (b) Portion of ϕ_i within the kernel functions.

on the average, no more than 10% of the kernel $k_i(r)$ is orthogonal to the basis functions. Figure 7b shows that about 50% of the function space of the basis functions is outside the kernel function space. Thus, the basis functions contain most of the information within the kernels and also add additional information about the anticipated aerosol size distributions.

The approximation of $f(r_0)$ as a linear sum of the basis functions results in some uncertainty as to the location of the radii r_0 in the number density $f(r_0)$. This uncertainty can be written as

$$f(r_0 + \Delta r_0) = \sum_{i=1}^{p} a_i \phi_i(r_0) \tag{81}$$

where Δr_0 is the uncertainty in the location r_0 for the number density at r_0. The magnitude of the uncertainty Δr_0 can be estimated by substituting a very narrow function at r_0 for $n(r)$ in Eq. (72) and constructing a number density function [Eq. (81)]. The limit of a very narrow function is a delta function. In numerical form, the delta function will be approximated by a normal distribution with a standard deviation of 0.01 μm.

The results of constructing 10 functions [Eq. (81)] from 10 delta functions centered at various locations r_0 are given in Figure 8, where it is shown that the uncertainty Δr_0 is about 0.25 μm, and the locations r_0 are exactly at the locations of the original delta functions. Figure 8a shows one of the constructed delta functions of Figure 8.

In order to estimate the error in deducing average number density in radii interval Δr=0.5 μm, a weighting function $w(r)$ was constructed from the kernels which contain the additional information from the

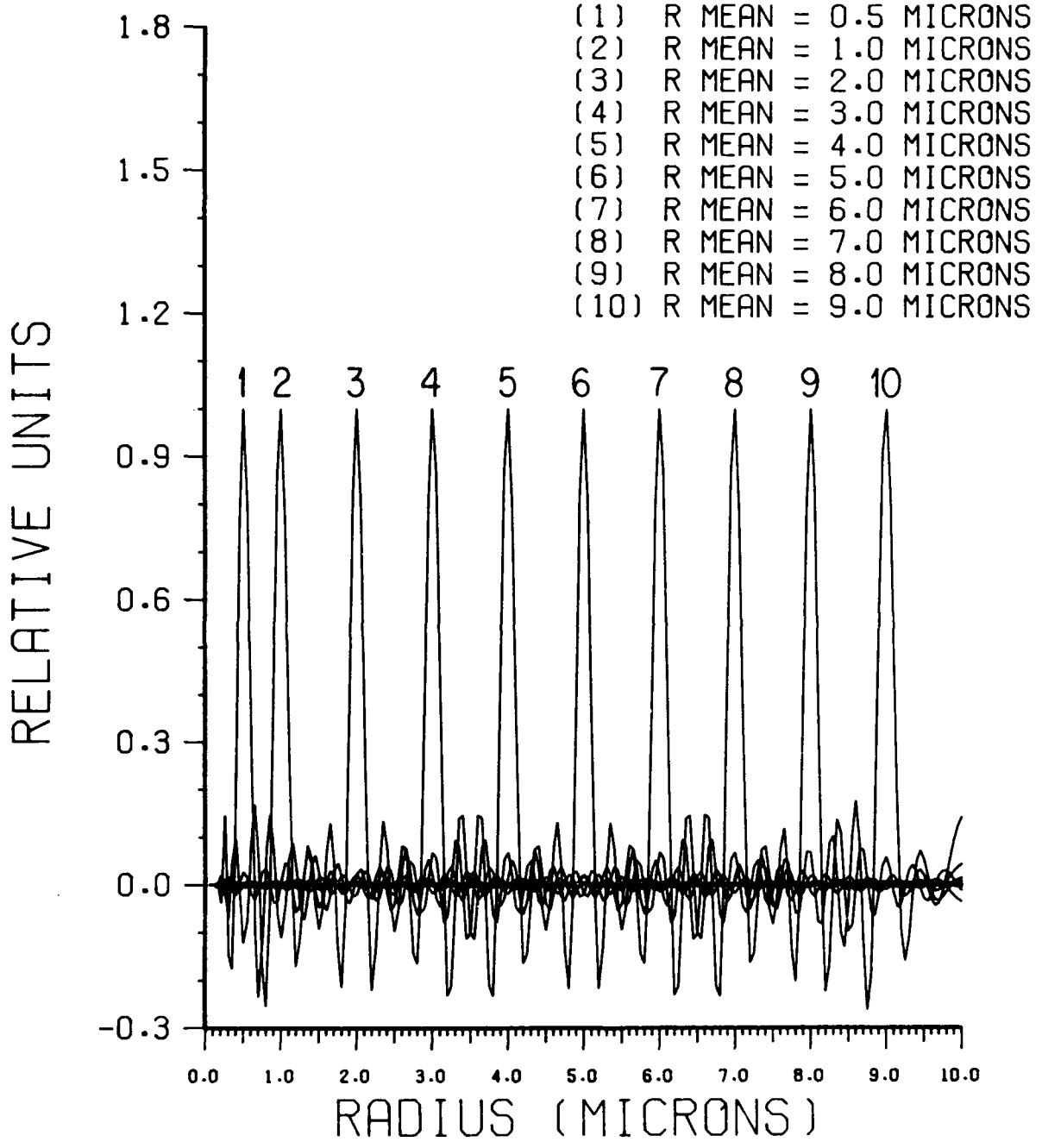


Figure 8. Approximation of delta functions by using the "pseudo"-empirical basis functions.

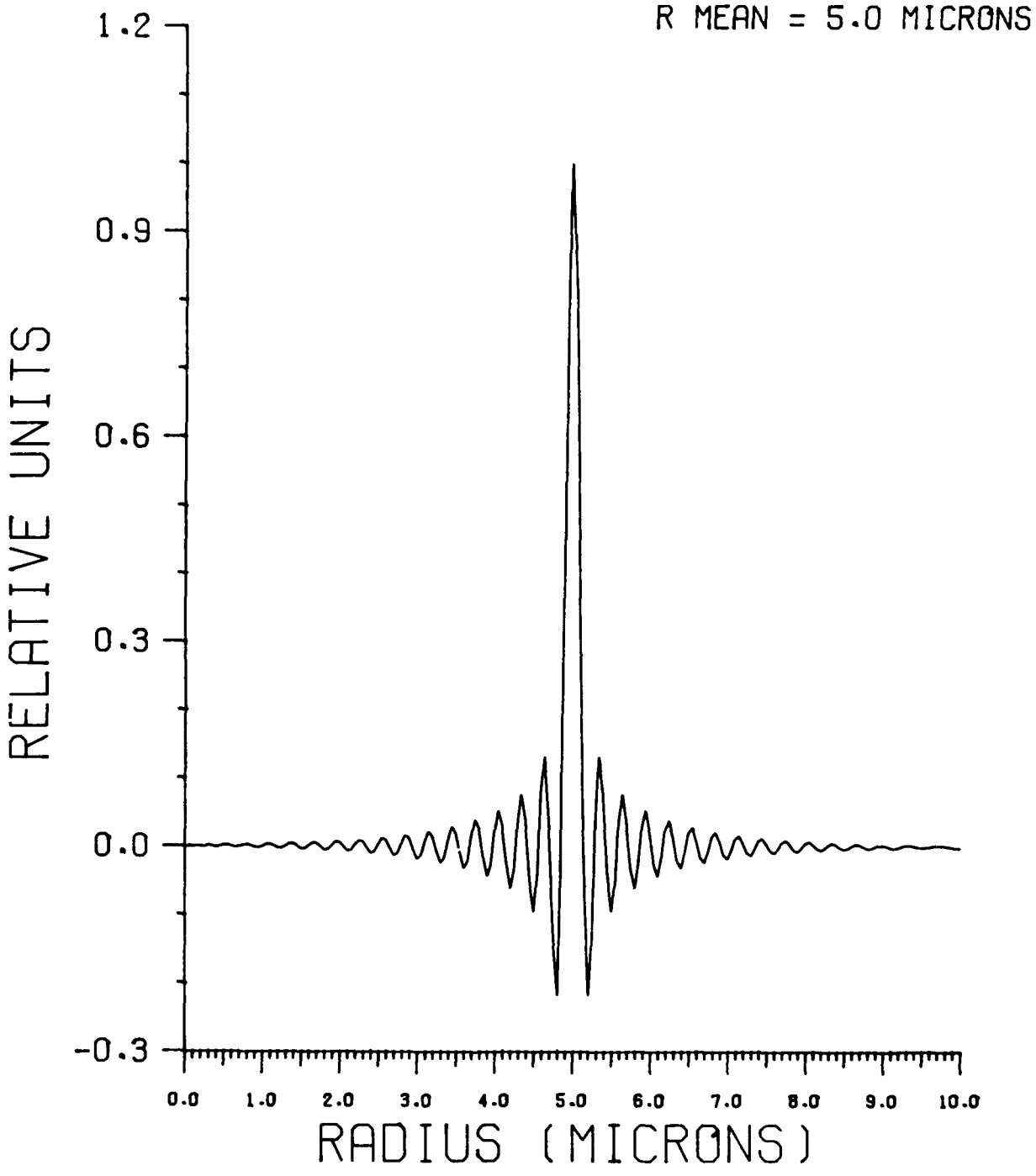


Figure 8a. Enlargement of approximation of delta function 6 of Fig. 8 using the "pseudo"-empirical basis functions.

pseudo-empirical functions $[B_i(x)]$.[29, Eq. 54] This error was calculated to be 50%. In Section 2.5, it will be shown (Fig. 9) that this estimate is a realistic estimate, and results of inversions performed with pseudo-empirical basis functions will be reported.

2.5 Results

Simulated measurements of backscattered radiation and refractive indices were provided for 15 wavelengths (between 0.2 μm and 9 μm) and for four aerosol size distributions (denoted by data sets G, H, I, and J). (Dr. J. Bottinger, private communication) These measurements were perturbed with various random errors which were normally distributed with a standard deviation of 10%. In this section, inversions for these four data sets will be performed along with analyses of the inversion results.

Figures 9a-d show the results of the inversion for data sets G, H, I, and J. In these figures, pseudo-empirical orthogonal functions were used to obtain two solutions for each data set. The first solution was obtained using the positivity constraint only ($\gamma_S = 0$). This solution is denoted SOL.1. For the second solution (SOL.2), both the positivity and the smoothing constraints were used. The dashed line (no symbols) represents the true solution for each of the four data sets. Figure 10 shows the provided simulated measurements for data G and the computed measurements from SOL.1 and SOL.2 from Figure 9a, all of which are a function of wavelength. The other solutions shown in Figure 9b-d reproduced the simulated measurements equally well and are not shown here. Figure 11 shows the iterative process used in solving Eq. (68) for data set I. The residual error (the difference

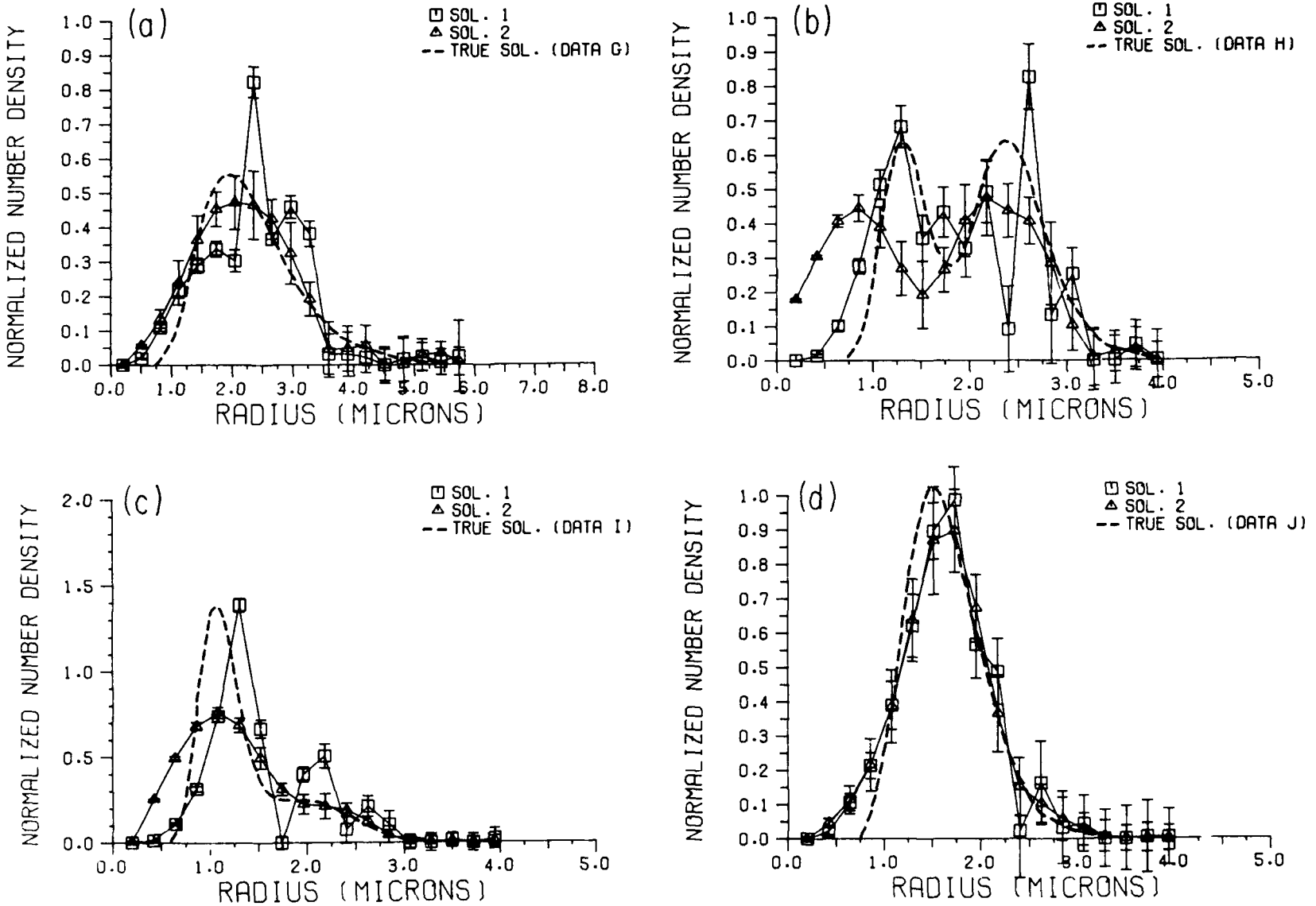


Figure 9. Inversion results for data sets G, H, I, and J, using the positivity constraint only (SOL.1) and a combination of positivity and smoothing constraints (SOL.2): (a) for data G, (b) for data H, (c) for data I, and (d) for data J.

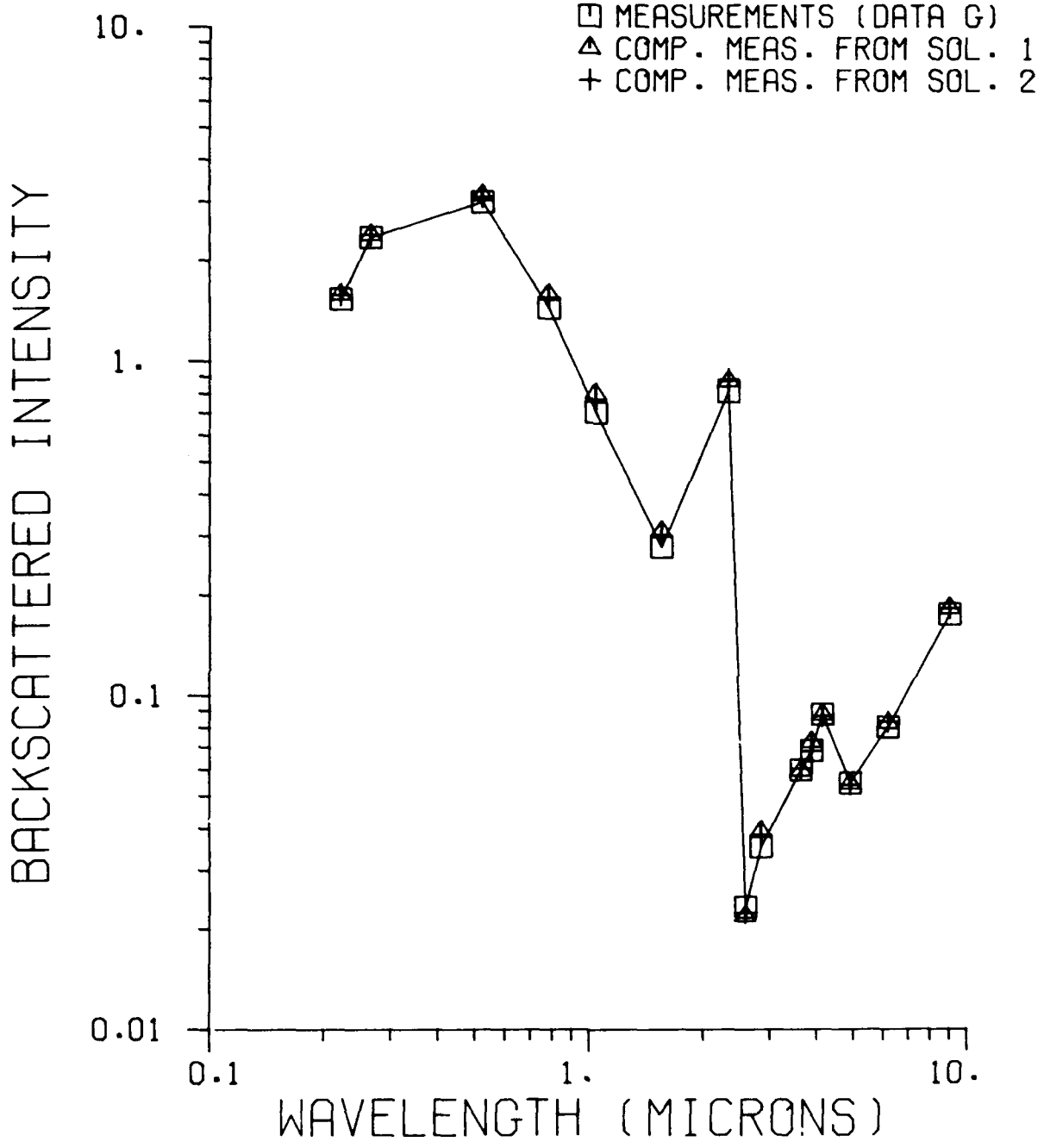


Figure 10. Computed measurements from inversions' solutions for data G and the input measurements.

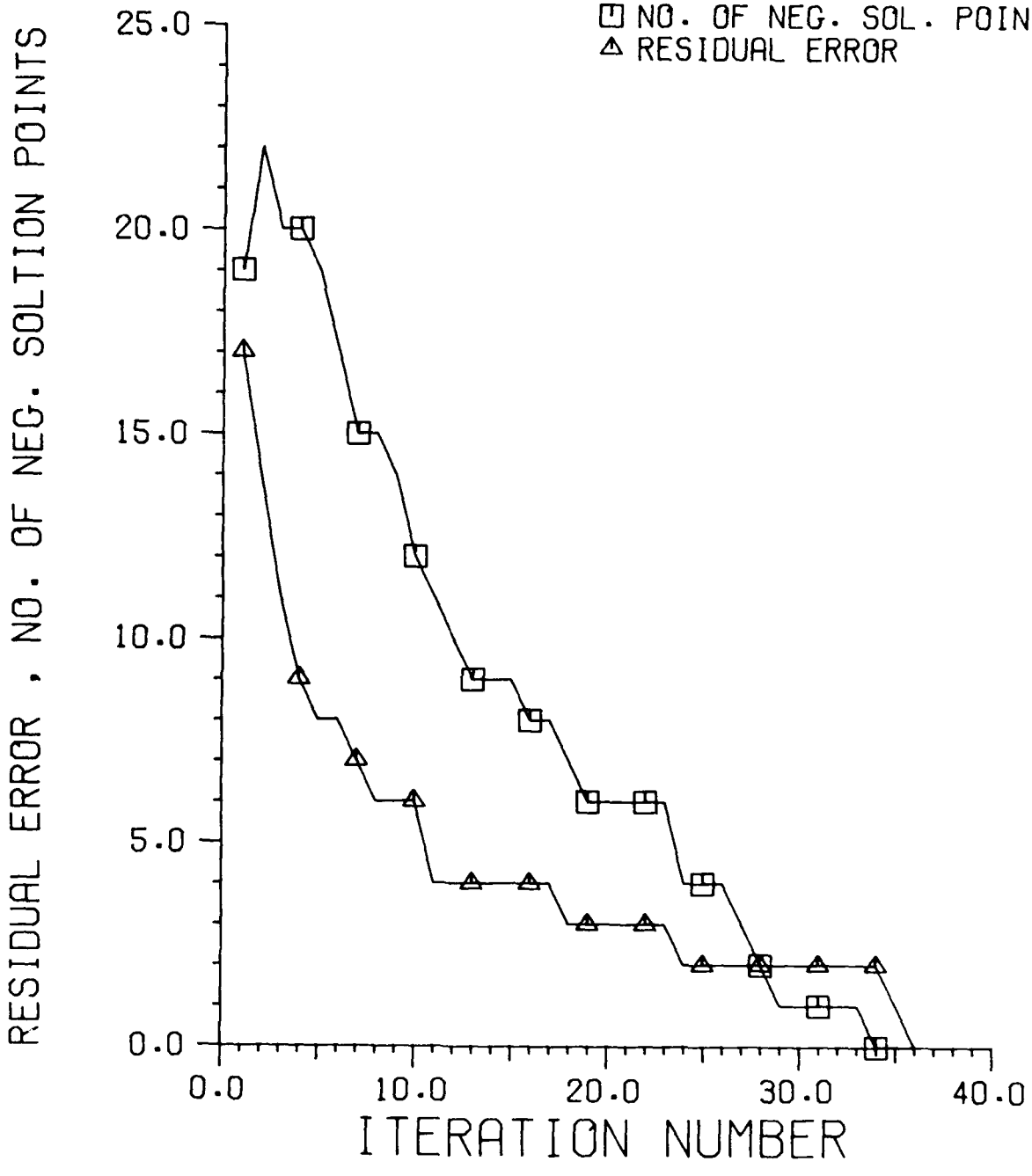


Figure 11. Residual error and number of negative solution points vs. iteration number during the iteration process for inversion of data "I".

between the simulated measurements and the computed measurements from the solution $\vec{f}$) and the number of negative solution points converge toward zero as the iteration number increases. The residual error is considered to be zero when the computed measurements are within 10% of the simulated measurements. (The simulated measurements contained 10% random errors.)

An inspection of the solutions presented in Figure 9 shows that SOL.1 differs greatly from SOL.2, but both can represent a real aerosol size distribution (*i.e.*, no negative aerosols), and both reproduce the simulated measurements equally well. SOL.2 contains a smoothing constraint on the unknown solution. An inspection of SOL.2 with reference to the true solution in Figure 9 reveals the assumption of smoothness. In Figure 9a, SOL.2 describes very well the true solution, which is smooth. In Figure 9b, the smoothing constraint forces SOL.2 to be much wider than the true solution. In Figure 9c, it may be noted that, for radii less than 1.5 μm, the smoothing constraint does not allow SOL.2 to reproduce the narrow peak at 1.0 μm. For radii larger than 1.5 μm, the true solution is nearly a straight line for which the second derivative is zero. In this part of the curve, the assumption of smoothness is accurate and is reflected in SOL.2. SOL.1, which was obtained by assuming only that $f(r) > 0$, preserves the width of the true solutions, as can be seen in Figures 9b-d. This type of comparison between the inversion solution and the true solution is possible only for simulated measurements. If both solutions, SOL.1 and SOL.2, were obtained from real measurements, both would be equally plausible.

The aerosol size distribution is often an input parameter in radiative transfer models and climate models and in other models for

atmospheric physics. The optical properties which are of interest for these models are: total scattering, absorption and extinction cross sections, single scattering albedo. and phase function of the aerosol size distribution. Figure 12a shows the ratio of computed optical properties from SOL.1 and SOL.2 for data G as a function of wavelength. From the figure, it can be seen that the difference between the computed optical properties from SOL.1 and SOL.2 (in Fig. 9a) is less than 10%. Figure 12b shows the ratio of computed phase function for three wavelengths from SOL.1 and SOL.2 for data G. The phase functions computed from SOL.1 and SOL.2 differ by less than 20%. Results of computation of ratios for the same optical properties for solutions for data sets H, I, and J are similar to Figures 12a and 12b and are not shown.

Physical properties such as average radius, total mass, total area, and average number density were computed for all solutions presented in Figures 9a-d. The ratio of computed physical properties from SOL.1 and SOL.2 for each data set is shown in Figure 13. The difference between computed physical properties for SOL.1 and SOL.2 is less than 20%. Therefore, either solution (SOL.1 or SOL.2) can be used equally well as an input parameter for calculating integrated optical and physical properties of the aerosol size distribution.

2.6 Summary

The pseudo-empirical orthogonal method of solution[29] was applied for determining aerosol size distributions from backscattered measurements. Two types of constraints were employed: a positivity constraint alone and a combination of positivity and smoothing constraints. The positivity constraint can be useful when the aerosol size distribution is known to be a narrow distribution or an unsmooth function. The use

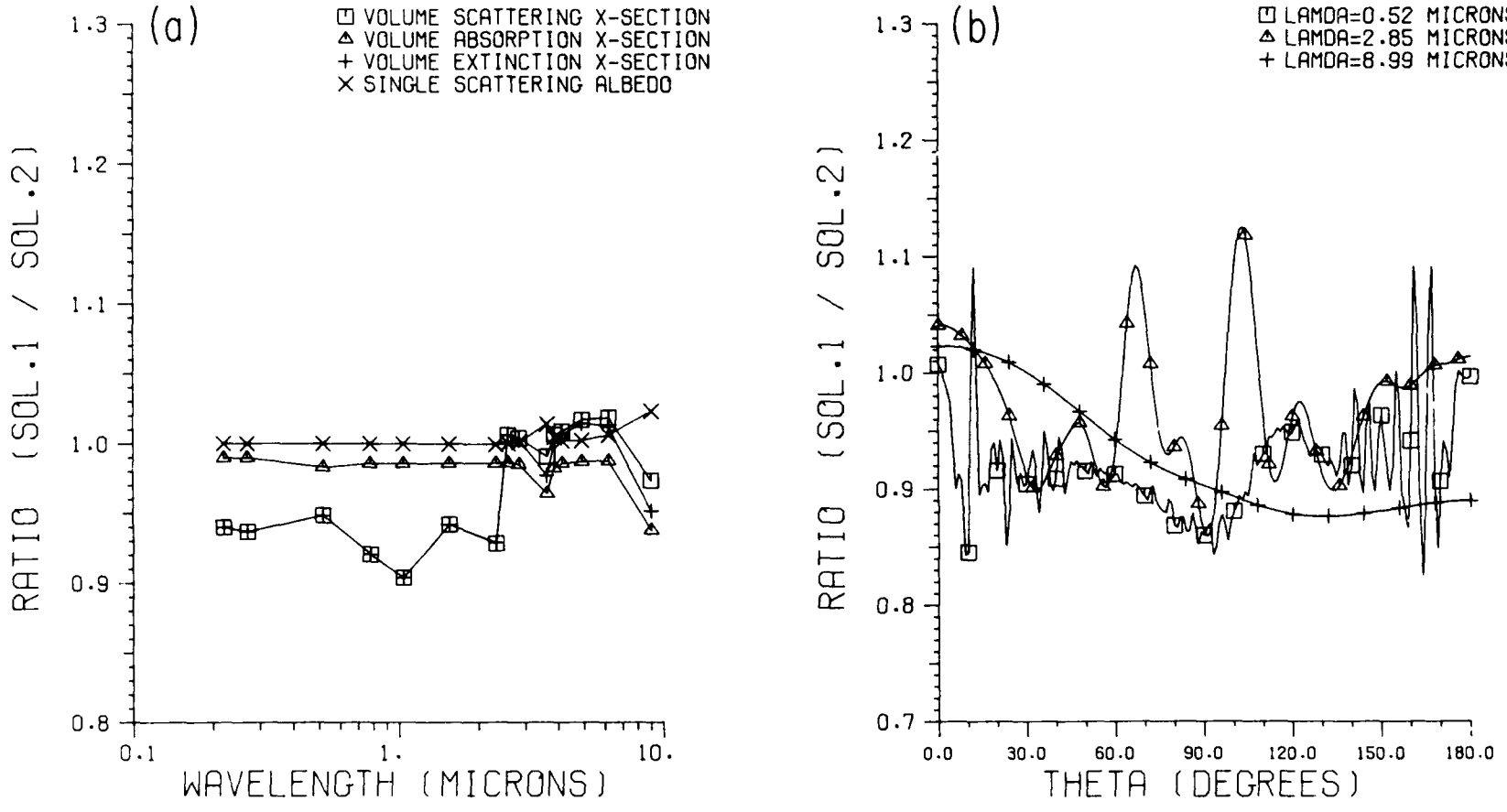


Figure 12. Ratio of computed optical properties from SOL.1 and SOL.2 for data G: (a) Volume, absorption and extinction cross-sections, single-scattering albedo, and (b) phase functions for three wavelengths.

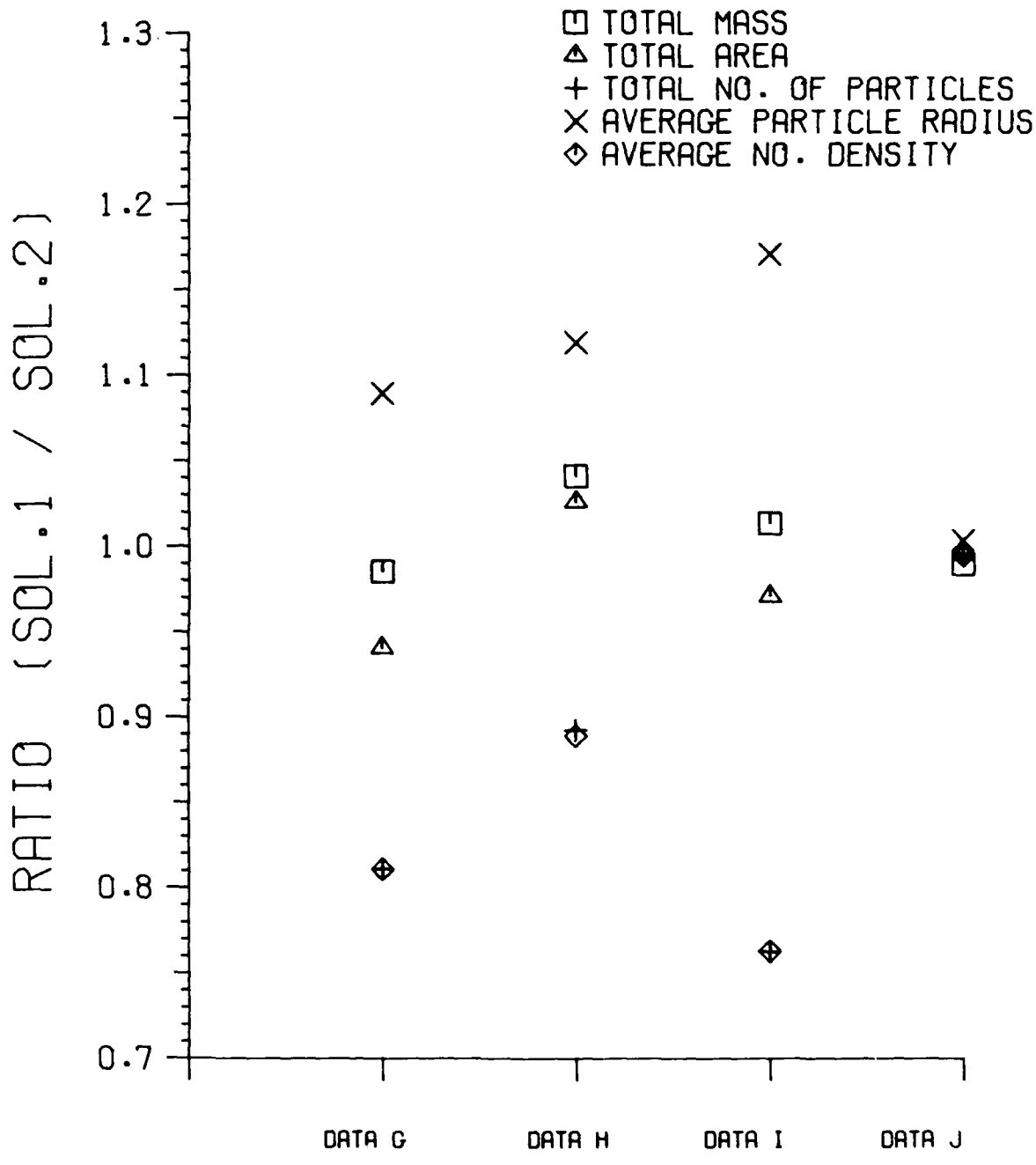


Figure 13. Ratio of physical properties computed from SOL.1 and SOL.2 for data sets G, H, I and J.

of the positivity constraint allows for a solution of Eq. (68) by an iterative process which converges very quickly (after 40 iterations).

The basis functions constructed from the "pseudo"-empirical functions can reproduce many types of aerosol distributions and the resulting measurements with good accuracy. The use of these basis functions does not overly restrict the variety of possible physical aerosol size distribution solutions. However, the use of empirical basis functions is a constraint on the solution, inasmuch as it makes it possible to obtain aerosol size distributions from the backscattering kernels, which otherwise would not be possible. The constraints (*i.e.*, smoothing, empirical basis functions) produce a solution which is no longer completely objective but which reflects the assumptions that are built into the constraints.

The accuracy of the solution [f(r)] at a discrete value of the radius is about ± 50%. Therefore, interpretation of features of the solution for specific radii can lead to wrong conclusions. Features in the solution with widths smaller than 0.3 μm in radius cannot be resolved. The inferred aerosol size distribution solution can be used for calculating integrated optical and mechanical properties of the particulates. The quality of the solution depends on the applicability of the constraints for the given problem. However, the inverse problem cannot be solved without using some constraints.

If no constraints are used and if the problems of instability in the inverse process are ignored, the types of solutions and the information content which can be obtained from the backscattered measurements will be apparent in the analysis of the natural basis function of the kernels. The backscattering kernels contain more information about

aerosol size distribution than do most other optical kernels. However, even if 40 optimal kernels with an accuracy of 0.5% are used for the inversion process, solutions such as Junge-type distributions and log-normal distributions, which are known to be present in the atmosphere cannot be obtained. The error in inferring size distributions at discrete radii is about ± 80%.

The volume backscattering cross-sections (to be used as the measurement vector $\vec{g}$) can be inferred from the solution to the lidar equation with an accuracy of no better than 10-15%.[43] In cases where the lidar measurements are restricted to relatively short ranges, volume backscattering cross-sections can be inferred with much better accuracy. However, the solution, $f(r)$, continues to be mathematically and physically non-unique. The physical non-uniqueness is caused by the uncertainties about particulate refractive indices and particle shapes. Different combinations of imaginary refractive indices, possible particle shapes and orientations, and various breadths of aerosol size distributions can all produce similar backscattering measurements. The resulting aerosol size distribution solution is, therefore, a plausible but non-unique solution for the inversion of multi-wavelength backscattered radiation, or any other type of optical measurements.

In this work, the measurements used contained 10% random error, a magnitude of error typical of atmospheric measurements. This error is, of course, reflected in the results (*i.e.*, 20% deviation in properties calculated from SOL.1 and SOL.2). It may be noted that the deviation between the derived solutions and the true solution can be much less in a laboratory experiment where the accuracy in the measurements can be improved and the refractive index and the sphericity of the particles can be controlled.

2.7 Literature Cited

1. J. M. Prospero, R. J. Charlson, V. Mohnen, R. Jaenicke, A. C. Delany, J. Moyers, W. Zoller, and K. Rahn, "The Atmospheric Aerosol System: An Overview," in *Reviews of Geophysics and Space Physics, Vol. 21* (American Geophysical Union, Washington, D.C., 1983) pp. 1607-1629.

2. B. M. Herman, S. R. Browning, and R. J. Curran, "The Effect of Atmospheric Aerosols on Scattered Sunlight," J. Atmos. Sci. **28**, 419 (1971).

3. O. B. Toon and J. B. Pollack, "A Global Average Model of Atmospheric aerosols for radiative transfer calculations," J. Appl. Meteorol. **15**, 225 (1976).

4. J. V. Dave, "Determination of Size Distribution of Spherical Polydispersion Using Scattered Radiation Data," Appl. Opt. **10**, 2035 (1971).

5. M. D. King, D. M. Ryne, B. M. Herman, and J. A. Reagan, "Aerosol Size Distribution Obtained by Inversion Optical Depth Measurements," J. Atmos. Sci. **35**, 2153 (1978).

6. P. T. Walter, "Practical Applications of Inverting Spectral Turbidity Data to Provide Aerosol Size Distribution," Appl. Opt. **19**, 2353 (1980).

7. J. H. Heintzenberg, H. Quenzel, and E. Thomella, "Information Content of Optical Data with Respect to Aerosol Properties: Numerical Studies with a Randomized Minimization-Search Technique Inversion Algorithm," Appl. Opt. **20**, 1308 (1981).

8. E. E. Uthe, "Particle Size Evaluation Using Multiwavelength Extinction Measurements," Appl. Opt. **21**, 454 (1982).

9. E. Trakhovsky, S. G. Lipson, and A. D. Devir, "Atmospheric Aerosol Investigated by Inversion of Experimental Transmittance Data," Appl. Opt. **21**, 3005 (1982).

10. M. D. King, "Sensitivity of Constrained Linear Inversion to the Selection of the Lagrange Multiplier," J. Atmos. Sci. **39**, 1355 (1982).

11. G. K. Yue and A. Deepak, "Retrieval of Stratospheric Aerosol Size Distribution from Atmospheric Extinction of Solar Radiation of Two Wavelengths," Appl. Opt. **22**, 1639 (1983).

12. J. T. Twitty, "The Inversion of Aureole Measurements to Derive Aerosol Size Distribution," J. Atmos. Sci. **32**, 584 (1979).

13. M. A. Box and A. Deepak, "Retrieval of Aerosol Size Distribution by Inversion of Simulated Aureole Data in the Presence of Multiple Scattering," Appl. Opt. **18**, 1376 (1979).

14. A. Deepak and M. A. Box, "Representation of Aerosol Size Distribution Data by Analytic Models," in *Atmospheric Aerosols: Their Formulation, Optical Properties, and Effects*, A. Deepak, Ed. (Spectrum, Hampton, Virginia, 1982), pp. 79-109.

15. N. T. O'Neill and J. R. Miller, "Constrained Linear Inversion of Optical Scattering Data for Particle Size Spectra: An Approach to Angular Optimization," Appl. Opt. **21**, 1231 (1982).

16. R. Santer and M. Herman, "Particle Size Distributions From Forward Scattered Light Using Chahine Inversion Scheme," Appl. Opt. **22**, 2294 (1983).

17. E. Trakhovsky and U. P. Oppenheim, "Determination of Aerosol Size Distribution from Observation of the Aureole Around a Point Source: 1. Theoretical," Appl. Opt. **23**, 1003 (1984).

18. G. E. Shaw, "Inversion of Optical Scattering and Spectral Extinction Measurements to Recover Aerosol Size Spectra," Appl. Opt. **19**, 988 (1979).

19. N. T. O'Neill and J. R. Miller, "Combined Solar Aureole and Solar Beam Extinction Measurements: 2. Studies of the Inferred Aerosol Size Distributions," Appl. Opt. **23**, 3697 (1984).

20. J. Heintzenberg and R. M. Welch, "Retrieval of Aerosol Size Distribution from Angular Scattering Functions: Effects of Particles, Composition, and Shape," Appl. Opt. **21**, 822 (1982).

21. C. D. Capps, R. L. Henning, and G. M. Hess, "Analytic Inversion of Remote Sensing," Appl. Opt. **21**, 3581 (1982).

22. V. E. Zuev and I. E. Natts, *Inverse Problems of Lidar Sensing of the Atmosphere* (Springer-Verlag, Berlin, 1983).

23. A. Ben-David and B. M. Herman, "Method for Determining Particle Size Distribution by Non-Linear Inversion of Backscattered Radiation," Appl. Opt. **24**, 1037 (1985).

24. D. Deirmendjian, "A Survey of Light Scattering Techniques Used in the Remote Monitoring of Atmospheric Aerosols," in *Reviews of Geophysics and Space Physics, Vol. 18* (American Geophysical Union, Washington, D. C., 1980), pp. 341-360.

25. A. I. Carswell, "Lidar Measurement of the Atmosphere," Canadian J. of Physics **61**, 378 (1983).

26. S. Twomey and H. B. Howell, "Some Aspects of the Optical Estimation of Microstructures in Fog and Cloud," Appl. Opt. **6**, 2125 (1967).

27. J. A. Reagan and B. M. Herman, "Three Optical Methods for Remotely Measuring Aerosol Size Distributions," AIAA J. **10**, 1401 (1972).

28. G. S. Kent, K. Glenn, U. Yue, O. Farrukh, and A. Deepak, "Modeling Atmospheric Aerosol Backscatter at CO_2 Laser Wavelengths: 1. Aerosol Properties, Modeling Techniques, and Associated Problems," Appl. Opt. **22**, 1655 (1983).

29. A. Ben-David and B. M. Herman, "The Inverse Problem and the Pseudo-Empirical Orthogonal Function Method of Solution. Part I: Theory," Appl. Opt., this issue (1987).

30. S. Twomey, "Information Content and Indirect Sensing Measurements," J. Atmos. Sci. **27**, 515 (1970).

31. S. Twomey, "Information Content in Remote Sensing," Appl. Opt. **13**, 942 (1974).

32. S. Twomey, *Introduction to the Mathematics of Inversion in Remote Sensing and Indirect Measurements* (Elsevier, New York, 1977), 243 p.

33. E. M. Patterson and D. A. Gillette, "Commonalities in Measured Size Distributions for Aerosols Having a Soil Derived Component," J. Geophys. Res. **82**, 2074 (1977).

34. E. M. Patterson, "Size Distributions, Concentrations, and Composition of Continental and Marine Aerosols," in *Atmospheric Aerosols: Their Formulation, Optical Properties, and Effects*, A. Deepak, Ed. (Spectrum, Hampton, Virginia, 1982), pp. 1-39.

35. K. Willeke and T. Whitby, "Atmospheric Aerosols, Size Distribution Interpretation," J. Air Poll. Cont. Assoc. **25**, 529 (1975).

36. C. E. Junge, "Atmospheric Chemistry," Advances in Geophysics **4**, 1 (1958).

37. J. W. Fitzgerald, "Effect of Relative Humidity on the Aerosol Backscattering Coefficient at 0.694 and 10.6 μm Wavelengths," Appl. Opt. **23**, 411 (1984).

38. E. J. McCartney, *Optics of the Atmosphere* (John Wiley & Sons, New York, 1976), 408 p.

39. J. A. Reagan, M. V. Apte, T. V. Bruhns, and O. Youngbluth, "Lidar and Balloon-Borne Cascade Impactor Measurements of Aerosols: A Case Study," Aerosol Sci. and Tech. **3**, 259 (1984).

40. S. Twomey, "On the Numerical Solution of Fredholm Integral Equations of the First Kind by the Inversion of the Linear System Produced by Quadrature," J. Assoc. Comput. Mach. **10**, 97 (1963).

41. A. Deepak and M. A. Box, "Experimental Validation of the Solar Aureole Technique for Determining Aerosol Size Distribution," Appl. Opt. **21**, 2236 (1982).

42. D. Deirmendjian, *Electromagnetic Scattering on Spherical Polydispersions* (American Elsevier, New York, 1969), 240 p.

43. P. B. Russell, T. J. Swissler, and M. P. McCormick, Methodology for Error Analysis and Simulation of Lidar Aerosol Measurements," Appl. Opt. **18**, 3783 (1979).